Proactive Caching for Low Access-Delay Services under Uncertain Predictions

RAN LIU, Northeastern University, USA
EDMUND YEH, Northeastern University, USA
ATILLA ERYILMAZ, Ohio State University, USA

Network traffic of delay-sensitive services has become a dominant part in the network. Proactive caching with the aid of predictive information has been proposed as a promising method to enhance the delay performance, which is one of the principal concerns of such services. In this paper, we analytically investigate the problem of how to efficiently utilize uncertain predictive information to design proactive caching strategies with provably good access-delay characteristics. First, we derive an upper bound for the average amount of proactive service per request that the system can support. Then we analyze the behavior of a family of threshold-based proactive strategies with a Markov chain, which shows that the average amount of proactive service per request can be maximized by properly selecting the threshold. Finally, we propose the UNIFORM strategy, which is the threshold-based strategy with the optimal threshold, and show that it outperforms the commonly used Earliest-Deadline-First (EDF) type proactive strategies in terms of delay. We perform extensive numerical experiments to demonstrate the influence of thresholds on delay performance under the threshold-based strategies, and specifically compare the EDF strategy and the UNIFORM strategy to verify our results.

CCS Concepts: • **Mathematics of computing** → **Queueing theory**; **Stochastic processes**; *Mathematical optimization*; • **Networks** → *Packet scheduling*; *Network performance analysis*;

Keywords: Proactive Caching; Prefetching; Queueing Theory; Markov Chain

ACM Reference Format:
Ran Liu, Edmund Yeh, and Atilla Eryilmaz. 2019. Proactive Caching for Low Access-Delay Services under Uncertain Predictions. In *Proc. ACM Meas. Anal. Comput. Syst.*, Vol. 3, 1, Article 2 (March 2019). ACM, New York, NY. 46 pages. https://doi.org/10.1145/3311073

1 INTRODUCTION

The traffic load in the network has been growing dramatically in recent years. Among all types of traffic in the network, delay-sensitive traffic, such as video, gaming, virtual reality (VR) and augmented reality (AR), has been a dominant component. According to a report from Cisco [7], video traffic takes up 73% of all the IP traffic in 2016 and is forecasted to be 82 % by 2021; Internet gaming traffic will grow nearly tenfold from 2016 to 2021; and the VR and AR traffic will increase 20-fold from 2016 to 2021. The delay performance of delay-sensitive services has a great impact on the revenue of companies like Amazon and Google[12]. Therefore, it is crucial to improve the delay performance of delay-sensitive services in communication networks.

Authors' addresses: Ran Liu, Electrical and Computer Engineering, Northeastern University, 460 ISEC Building, 805 Columbus Avenue, Boston, MA, 02115, USA, rliu1@ece.neu.edu; Edmund Yeh, Electrical and Computer Engineering, Northeastern University, 413 ISEC Building, 805 Columbus Avenue, Boston, MA, 02115, USA, eyeh@ece.neu.edu; Atilla Eryilmaz, Electrical and Computer Engineering, Ohio State University, 708 Dreese Laboratories, 015 Neil Avenue, Columbus, OH, 43210, USA, eryilmaz.2@osu.edu.

2476-1249/2019/3-ART2 $15.00
https://doi.org/10.1145/3311073

Distributed caching techniques are seen as an effective method to achieve this goal, and there has been extensive work in this area, such as [21],[10],[13]. Caching networks can reduce a considerable amount of traffic by caching data objects locally, and thereby greatly reduce the time and network resources to fetch the requested data object from the server.

Proactive caching techniques, which take advantage of predictive information of user requests and network states, utilize the spare bandwidth resources and potentially place the data objects in the caches before requests are generated. In [4], two experimental cases were carried out to show the promise of proactive caching in 5G wireless networks. There has also been considerable literature on prediction methods based on user behaviors (e.g., [16],[1]), which showed certain predictability of user demands. However, these work did not reveal the fundamental insights on how much improvement in system performance we can expect by utilizing prediction information.

There has been some recent analytical work on studying the fundamental benefits achieved from predictive information and proactive scheduling in networks. In [19], the authors characterized the diversity and multicasting gains of proactive caching using large-deviations theory under the assumption of perfect predictions. In [15] and [18], the authors studied a cost optimization problem in a multi-user single-server system with proactive scheduling. The authors proposed a model with uncertainties in user demands and channel states, and designed a proactive scheduling algorithm, which was proved to be asymptotically optimal in cost. In [2], the authors considered a profit maximization problem for a carrier and a cost minimization problem for users with predictive information of user demands. In [9], the authors studied the delay performance of a backpressure algorithm in a downlink system with perfect predictions, where the requested objects and corresponding request epochs are accurately predicted. The authors proved that the average queueing delay asymptotically goes to 0 as the prediction window size goes to infinity. They also analyzed the impact of prediction window size on the delay performance. Following this work, the authors of [22] studied the fundamental queueing performance of a single queue proactive system. They analyzed a variety of scenarios with different arrival and service processes, different prediction window sizes, and different types of imperfect predictions. They showed that proactive services exponentially reduce delay, especially in a lightly-loaded case. A related work [6] designed and analyzed a predictive scheduling algorithm which maximizes the timely-throughput, which is the total traffic received before the deadlines. All the work mentioned above shows that taking advantage of predictive information greatly improves the system performance.

Our work aims to study the characteristics of proactive caching based on uncertain predictive information from a fundamental queueing theory perspective. Different from the work of [22], we not only look at the basic queueing dynamics of the proactive system but also further explore how to strategically utilize uncertain predictions to enhance delay performance. In terms of delay performance, we take the Earliest-Deadline-First (EDF) type strategy, which has been widely used in network scheduling problems, as a competitive baseline in our analysis. There have been many work (e.g., [3],[17] and [11]) which studied the delay performance of the EDF strategy. In the proactive caching context, we consider the 'deadlines' to be the predicted arrival epochs. The authors of [9] has proved that the EDF strategy achieves optimal delay performance under perfect predictions.

The main contributions and the structure of this paper are listed as follows:

- We propose a request model which characterizes the request uncertainty by introducing a *potential request process*. We aim to maximize the average amount of proactive service for each request. We introduce our system model and problem formulation in Section 2.
- Based on the request-model with uncertainty, we reveal the iterative nature of bandwidth resource assignment between reactive service and proactive service, by comparing the EDF

strategy with a First-Come-First-Serve reactive strategy as an example. As a result, we derive an upper bound on how much proactive service per request that the system can support. We discuss the comparisons and derive the bounds in Section 3.

- For the purpose of analysis, we define a family of threshold-based proactive strategies, where the threshold determines the maximal amount of proactive service to be done for each future potential request. We construct a Markov chain to analyze the asymptotic behaviors of the proactive system under the threshold-based strategies. We prove that the UNIFORM strategy, which is the threshold-based strategy with the optimal threshold, is the solution to the optimization problem we proposed. We obtain an important insight on how to design an optimal proactive strategy: the strategy should balance proactive service among the predictions in nearer future and farther future based on prediction uncertainties. We present the threshold-based strategies, the corresponding Markov chain, and the corresponding analysis in Section 4.
- We analytically compare the delay performance of the EDF type strategy with the UNIFORM strategy. Although one would intuitively expect the EDF strategy to achieve desirable delay performance based on its performance in previous network scenarios, we prove that the delay performance of the EDF strategy is always worse than the UNIFORM strategy in all the non-trivial cases. We show the analysis in Section 5.
- We conduct extensive range of experiments to show the delay performance of the threshold-based strategies with different thresholds. Specifically, we compare the delay performance of the UNIFORM strategy with the EDF strategy in multiple network scenarios, with the reactive scheme as a baseline. The results show that proactive caching not only greatly improves delay performance in lightly-loaded cases as concluded in [22], but also works exceedingly well in the heavily-loaded scenario with the UNIFORM strategy. We also carry out experiments to show the impact of prediction window size on the delay performance for practicality concerns. The UNIFORM strategy still shows excellent delay performance with simple modifications. We show the numerical results in Section 6.

2 SYSTEM MODEL

2.1 Network Model

We consider a system with one server providing delay-sensitive services to the user, as shown in Figure 1. The system operates in continuous time from time 0. The user receives service from the server at a constant rate of μ $bits/sec$.

Request Processes: Requests arrive at the server according to the processes shown in Figure 2. The requests request same-sized data objects of s bits. The *Potential Request Process* is a Poisson Process $\{P(t); t > 0\}$ with an overall arrival rate of λ, where the ith arrival, i.e. *Potential Request* i, requests object $r_i \in \mathbb{Z}^{+1}$ at time $t_i \in \mathbb{R}^+$, where $0 < t_1 < t_2 < \ldots$. The *Actual Request Process* $\{A(t); t > 0\}$ is a thinned version of $P(t)$ where each arrival on $P(t)$ is an arrival on $A(t)$ with probability p, independent of all other arrivals. Let $\{R_i; i = 1, 2, \ldots\}$ be IID $Bernoulli(p)$ indicator random variables where $R_i = 1$ if the ith arrival on $P(t)$ is an arrival on $A(t)$. Thus, $A(t)$ is a Poisson process with an average arrival rate λp. For convenience, we denote an actual request with its index in $P(t)$ instead of $A(t)$.

An important assumption we make is that every potential request requests a different object, i.e. $r_i \neq r_j, \forall i \neq j$, i.e., the catalog size is assumed to be infinite. This assumption is motivated by many practical problems, e.g. 1) prefetching problems, where each prefetched object is usually considered to be specific for one user request, 2) applications where data objects are highly dynamic, like live

[1] $\mathbb{Z}^+ = \{1, 2, 3, \ldots\}$ in this paper.

Fig. 1. Network Model

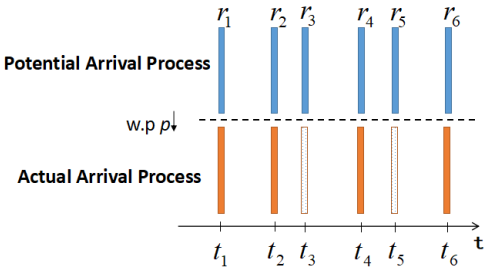

Fig. 2. Arrival Processes

streaming, online gaming, sensing data, cloud computing, etc., and 3) the small likelihood that a user would request for the objects that are recently requested. For more general applications, our aim is to (for simplicity) exclude the impact of popularity distributions and focus on the potential gains of proactive caching in the presence of uncertain predictive information.

Predictions: At time 0, the server knows the sequence of objects $(r_1, r_2, r_3, \ldots)$ to be requested by the arrivals in $\{P(t)\}$, and the probability p. It has no prior knowledge of the precise arrival epochs $\{t_i\}$, or the realizations of the indicator random variables $\{R_i\}$. The server observes $A(t)$ but not $P(t)$. At time $t > 0$, the sequence of indices for future potential requests from the server's viewpoint, or the *prediction window*[2], is:

$$\mathbf{\Pi}(t) = (I(t) + 1, I(t) + 2, I(t) + 3, \ldots) \tag{1}$$

where $I(t)$ is defined as:

$$I(t) \triangleq \max\{i | t_i < t, R_i = 1\} \tag{2}$$

i.e. the index of the most recent actual request before time t. The server proactively works on request i only if $i \in \mathbf{\Pi}(t)$ at time t.

The idea of this prediction model originates from perfect prediction models used in the work of [18], [9]. With our prediction model, we are able to tractably model uncertainties in whether potential requests are realized, as well as uncertainties in the request arrival epochs.

2.2 Service Model

In this section, we first describe the *reactive scheme*, where the server works only on requests made by actual request arrivals. We then introduce the *proactive schemes* where the server works on future potential requests when not serving requests made by actual requests.

Reactive Scheme: The server node serves only arrivals in the actual request process $A(t)$ based on strategy Ψ_R as described below. Upon observing an actual request i at time t_i, the object r_i is placed into the tail of a FIFO Queue with $V(t_i)$ of unfinished work, which is transmitted back to the user at rate μ, where $V(t)$ is defined as the total number of bits waiting to be transmitted in the queue at time t in the reactive scheme. If $V(t) = 0$, the system is *idle* at t.

Proactive Schemes: The server can proactively send a data object, partially or in entirety, to the user, which can store the data object in a local cache. Since our focus is on the effects of uncertain predictions, we assume for simplicity that the cache size is infinite.

[2]We assume the prediction window size to be infinite for simplicity of analysis. This assumption guarantees that the server always has predicted requests on which to do proactive work.

Let $U_i(t) \leq s$ be the proactive work done for request i by time t, i.e. the number of bits of object r_i sent to the user and stored in the cache by time t. Notice that for a request i, there is no reason to continue to proactively serve it after $t_{H(i)}$, where $H(i) \triangleq \min\left\{j \geq i | R_j = 1\right\}$ represents the first potential request after i which is realized. Let

$$U_i \triangleq \min\left\{s, U_i\left(t_{H(i)}\right)\right\}$$

be the total proactive work done for request i. For an actual request i ($R_i = 1$), $U_i = \min\{s, U_i(t_i)\}$. Define $S_i = s - U_i = \max\left\{0, s - U_i\left(t_{H(i)}\right)\right\}$ as the *reactive* part of object r_i which remains to be transmitted after the server stops proactively serving request i. For an actual request i, $S_i = s - U_i = \max\{0, s - U_i(t_i)\}$ bits need to be transmitted reactively at t_i. Let $\mathbf{U}(t) \triangleq \left(U_{I(t)+1}(t), U_{I(t)+2}(t), U_{I(t)+3}(t), \ldots\right)$ be the set of $U_i(t)$'s where $i \in \mathbf{\Pi}(t)$. At time t, based on $\mathbf{U}(t)$, the prediction window $\mathbf{\Pi}(t)$ and the queue size $V(t)$, a stationary proactive rate allocation strategy Ψ_P at the server is defined as:

$$\Psi\left(V(t), \mathbf{\Pi}(t), \mathbf{U}(t)\right) = \left\{\rho_V(t), \rho_{I(t)+1}(t), \rho_{I(t)+2}(t), \rho_{I(t)+3}(t), \ldots\right\}$$

where $\rho_V(t)$ is the rate allocated to serve the queue of $V(t)$, and $\rho_{I(t)+i}(t), i \geq 1$, is the rate allocated to fetch object $r_{I(t)+i}$ at time t. We assume that the data in $V(t)$ has higher priority than proactive traffic. That is, if $V(t) > 0$, then $\sum_{i \in \mathbb{Z}^+} \rho_{I(t)+i}(t) = 0$. Thus, we consider the set Γ_P of proactive strategies Ψ_P satisfying:

- (Reactive State) If $V(t) > 0$:

$$\rho_V(t) = \mu, \sum_{i=1}^{\infty} \rho_{I(t)+i} = 0 \tag{3}$$

- (Proactive State) If $V(t) = 0$:

$$\rho_V(t) = 0, \sum_{i=1}^{\infty} \rho_{I(t)+i} = \mu \tag{4}$$

- The limiting average amount of proactive work received per potential request

$$\overline{U} \triangleq \lim_{t \to \infty} \frac{\sum_{i=1}^{I(t)} U_i}{I(t)} \tag{5}$$

  exists for Ψ_P;
- The limiting average amount of proactive work received per actual request

$$\overline{U_A} \triangleq \lim_{t \to \infty} \frac{\sum_{i \in \mathbb{Z}^+ : i \leq I(t), R_i = 1} U_i}{A(t)} \tag{6}$$

  exists for Ψ_P.

An example of a strategy in Γ_P is the *Earliest-Deadline-First (EDF)* strategy. In the EDF strategy, if $V(t) = 0$ at time t, then $\rho_{J(t)} = \mu$, where $J(t) = \min\{i \in \mathbf{\Pi}(t) | U_i(t) < s\}$. We use EDF strategy as an important baseline policy throughout the paper for the purpose of analysis and comparisons. Given a sample path of arrival epochs and $\{R_i\}$ realizations, the evolutions of unfinished work in $V(t)$ under the EDF strategy and under the reactive scheme Ψ_R are compared, as shown in Figure 3.

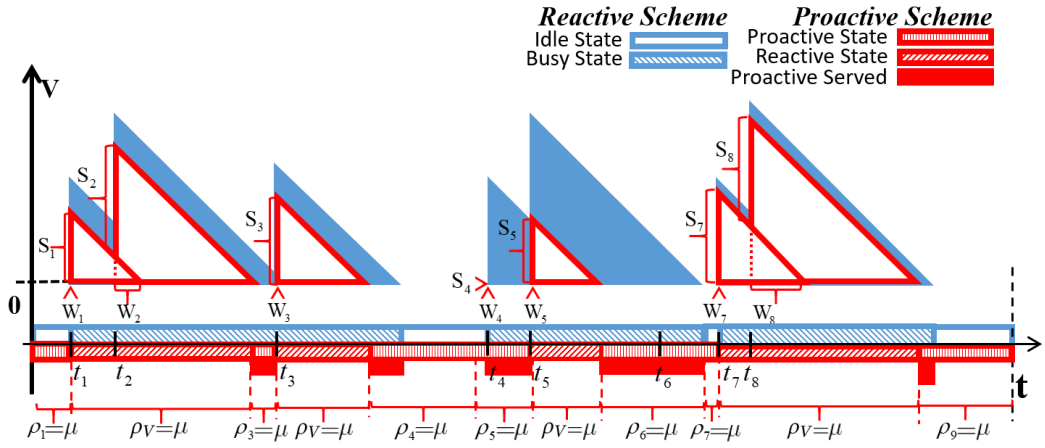

Fig. 3. The system runs from time 0 to t. Potential requests 1 to 8 arrive at t_1 to t_8 respectively during this period of time, with all potential requests realized except for request 6. The evolution of unfinished work $V(t)$ under the reactive scheme is plotted in blue and the evolution in the proactive scheme with the EDF strategy is plotted in red, with the corresponding states marked on the time axis. We also show the rate allocation in the proactive scheme.

2.3 Problem Formulation

As shown in Figure 3, there is less traffic served reactively in the proactive EDF scheme as compared with the reactive scheme. Reducing reactive traffic is doubly desirable since (1) the delay is reduced, and (2) there is more time for the server to do proactive work. Motivated by this, we study an optimization problem where the objective is to maximize the average amount of proactive work done for each request. Given λ, μ, p and s, our optimization problem then can be formulated as:

$$\begin{aligned} \underset{\Psi_P}{\text{maximize}} \quad & \overline{U}(\Psi_P) \\ \text{subject to} \quad & \Psi_P \in \Gamma_P \end{aligned} \tag{7}$$

where Γ_P is defined in (3)-(6). Let Ψ^* be an optimal solution to problem (7) and let $\overline{U}_{max} \triangleq \overline{U}(\Psi^*)$ denote the $\overline{U}$ achieved by Ψ^*. The solution to (7) is discussed and presented in Sections 3 and 4.

Operating Regimes: In fact, there is a limited region of λ we are interested in. In the region $0 \le \lambda < \frac{\mu}{s}$, $\overline{U}_{max} = s, w.p.1$ by Corollary 2 in [9] and Theorem 2 in [22]. With knowledge of $\mathbf{\Pi}(t)$, the server is able to proactively serve every request before its arrival epoch with probability 1, even if every request is realized. In the region $\lambda \ge \frac{\mu}{ps}$, the arrival rate of the actual request process is beyond the stability region of the network. According to [8], full knowledge of the future does not enlarge the stability region of the system. Thus, the queue $V(t)$ cannot be stabilized in this region. This implies that the server almost always works reactively, sparing no bandwidth for proactive service. In the region $\frac{\mu}{s} \le \lambda < \frac{\mu}{ps}$, an optimal solution Ψ^* to problem (7) is proposed and analyzed in Section 4. Thus, we have the following fact:

$$\overline{U}_{max} = \begin{cases} s, w.p.1 & if\ 0 \le \lambda < \frac{\mu}{s} \\ \overline{U}(\Psi^*), & if\ \frac{\mu}{s} \le \lambda < \frac{\mu}{ps} \\ 0, w.p.1 & if\ \frac{\mu}{ps} \le \lambda \end{cases} \tag{8}$$

Table of Notations	
$V(t)$	Unfinished reactive work at server node
s	Object size
μ	Constant service rate of the system
$P(t)$	Potential request process
$A(t)$	Actual request process
λ	Average arrival rate of $P(t)$
p	Probability that each potential request is realized
t_i	Arrival epoch of potential request i
R_i	Indicator random variable for whether request i is realized
U_i	Total amount of proactive service for request i
$U_i(t)$	Amount of proactive service for request i by time t
S_i	Amount of reactive work for request i
$\mathbf{\Pi}(t)$	Prediction window
$I(t)$	Index of the latest actual request before time t
$J(t)$	Index of the request to practively serve at t
$\overline{U}$	Limiting time average proactive service per potential request
$\overline{U_A}$	Limiting time average proactive service per actual request
U^*	Maximum limiting average proactive service per potential request
Ψ_P^{ϕ}	Threshold-based strategy with threshold ϕ
X_n	Markov chain
τ_n	Epoch of the nth transition

Table 1. Table of Notation

Delay Performance: The corresponding delay of Ψ^* is analyzed in Section 5. For a given $\Psi_P \in \Gamma_P$, we define the delay of an actual request i as

$$D_i = \begin{cases} \frac{V(t_i)}{\mu} + \frac{S_i}{\mu}, & if\ S_i > 0\ and\ R_i = 1 \\ 0, & otherwise \end{cases}$$

where $\frac{V(t_i)}{\mu}$ is the waiting time of object r_i in the queue at the server, and $\frac{S_i}{\mu}$ is the transmission time of the reactive part of object r_i. Define the limiting average delay per actual request as:

$$\overline{D} \triangleq \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+: t_i<t, R_i=1} D_i}{A(t)}$$

Denote the average delay per actual request under Ψ^* by $\overline{D}_{\Psi^*}$. We will derive the closed-form expression of $\overline{D}_{\Psi^*}$, and analytically demonstrate its advantage relative to average delay of the EDF proactive strategy.

3 RELATION BETWEEN REACTIVE SCHEME AND PROACTIVE SCHEMES

Proactive caching makes use of available link capacity when the system is idle (under the reactive scheme). A natural question to ask is how much proactive work can be done for each request on average. We can gain intuition from the example in Figure 3. First, the idle period in the reactive scheme can be utilized for proactive service. Then, by proactively serving actual requests (i.e., 1,2,3,4,5,7,8), reactive traffic is reduced so that available link capacity can be utilized more frequently for proactive service. This is indicated in Figure 3 by the intervals marked by solid red, named "Proactive Served". In the following, we study the characteristics of proactive service and derive an upper bound on $\overline{U}$.

Consider a set of sample paths corresponding to arrival epochs $\{t_i : i = 1, 2, \ldots\}$ and realizations $\{R_i = z_i : i = 1, 2, \ldots\}$ $(z_i \in \{0, 1\})$ under both the reactive scheme and a proactive scheme Ψ_P. We make the following definitions. The amount of time that $\Psi_P \in \Gamma_P$ works in the proactive state (namely Proactive Proactive) from 0 to t is:

$$T_{PP}(t) \triangleq |\{\tau \in (0, t] : V(\tau) = 0\}| \tag{9}$$

The amount of time that $\Psi_P \in \Gamma_P$ works in reactive state (namely Proactive Reactive) from 0 to t is:

$$T_{PR}(t) \triangleq |\{\tau \in (0,t] : V(\tau) > 0\}| \tag{10}$$

The limiting fraction of time that $\Psi_P \in \Gamma_P$ works in the reactive state and in the proactive state, respectively, are:

$$\alpha_{PR} \triangleq \lim_{t\to\infty} \frac{T_{PR}(t)}{t},\ \alpha_{PP} \triangleq \lim_{t\to\infty} \frac{T_{PP}(t)}{t} \tag{11}$$

Before we continue to study the relation between the reactive scheme and the proactive scheme, we first define two important properties of proactive strategies.

Definition 1 (Property 1 of Proactive Strategies). *A proactive strategy $\Psi_P \in \Gamma_P$ satisfies Property 1 if the following condition is satisfied:*

$$\lim_{t\to\infty} \frac{\sum_{i=I(t)+1}^{\infty} U_i(t)}{t} = 0, w.p.1 \tag{12}$$

The term $\sum_{i=I(t)+1}^{\infty} U_i(t)$ represents the total amount of proactive work done for potential requests in the prediction window $\mathbf{\Pi}(t)$ up to t. Although this part of proactive work may be requested eventually in the future, it does not contribute to the reduction of reactive work by time t. If $\sum_{i=I(t)+1}^{\infty} U_i(t)$ scales with t, it is likely that the corresponding $\overline{U}$ can be further improved by a strategy which invests more proactive service into requests in the near future. We will later formally analyze the influence of this property on our objective in Theorem 1.

Proposition 1. *For all $\Psi_P \in \Gamma_P$, we have*

$$\overline{U} \geq \overline{U_A},\ w.p.1$$

Proof. Please refer to Appendix A for the proof. □

We then have the following definition of the second property:

Definition 2 (Property 2 of Proactive Strategies). *A proactive strategy satisfies Property 2 if the following condition is satisfied:*

$$\overline{U_A} = \overline{U}, w.p.1 \tag{13}$$

Proposition 1 implies that in our setting, the average amount of proactive work per actual request is no more than the average amount of proactive work per potential request. On the other hand, it is more desirable that more proactive services are done for actual requests. With Property 1 and 2, we have the following theorem of proactive strategies.

Theorem 1. *Given μ, λ, s and p as system parameters, the limiting fractions of time that the server works in the proactive state and the reactive state, respectively, under $\Psi_P \in \Gamma_P$ satisfy*

$$\alpha_{PP} \leq \frac{\mu - \lambda ps}{\mu(1-p)}, w.p.1, \quad \alpha_{PR} \geq \frac{(\lambda s - \mu)p}{\mu(1-p)}, w.p.1$$

Equality holds in both inequalities if and only if the proactive strategy satisfies both Property 1 and Property 2.

Proof. Please refer to Appendix B for the proof. □

Theorem 1 implies that in order to maximize the fraction of time that the system works proactively, or equivalently minimize the fraction of time that the system works reactively, the proactive strategy Ψ_P must satisfy both Property 1 and Property 2. On the other hand, recall that we are interested in the operating regime $\frac{\mu}{s} \leq \lambda < \frac{\mu}{ps}$. If $\lambda = \frac{\mu}{s}$, we have $\alpha_{PR} = 0, w.p.1$ if and only if Ψ_P satisfies both

Property 1 and Property 2. If $\lambda = \frac{\mu}{ps}$, we have $\alpha_{PR} \geq 1, w.p.1$, which implies that the system almost always works reactively with any proactive strategy $\Psi_P \in \Gamma_P$. These results are consistent with previous discussions before (8).

Define

$$\overline{S} \triangleq \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\leq I(t)} S_i}{A(t)} \tag{14}$$

as the limiting average amount of reactive work of each actual request. Then based on Theorem 1, we have the following corollary on $\overline{U}$ and $\overline{S}$.

Corollary 1. *Given μ, λ, s and p satisfying $\frac{\mu}{s} \leq \lambda < \frac{\mu}{ps}$, the limiting average amount of proactive work per potential request under strategy $\Psi_P \in \Gamma_P$ satisfies*

$$\overline{U} \leq \frac{\mu - p\lambda s}{\lambda(1-p)} \triangleq U^*, w.p.1 \tag{15}$$

The limiting average amount of reactive work per actual request under strategy $\Psi_P \in \Gamma_P$ satisfies

$$\overline{S} \geq \frac{\lambda s - \mu}{\lambda(1-p)} \triangleq S^*, w.p.1 \tag{16}$$

where equality holds in both inequalities if and only if strategy Ψ_P satisfies both Property 1 and Property 2.

Proof. Please refer to Appendix C for the proof. □

Corollary 1 shows that the limiting average amount of proactive work done per potential request is maximized if and only if a proactive strategy Ψ_P satisfies both Property 1 and Property 2. By Property 2, $\overline{U_A}$ is maximized under the same condition. Therefore, the optimal solution to the objective in (7) should be proactive strategies which satisfy both Property 1 and Property 2. We will construct such a proactive strategy, and also explain why the EDF strategy is not an optimal solution in the next section.

4 THRESHOLD-BASED PROACTIVE STRATEGY AND MARKOV CHAIN

In order to construct an optimal proactive strategy to solve (7), we first define a family of threshold-based strategies in Γ_P. We then analyze the asymptotic behaviors of the threshold-based proactive strategies by constructing and analyzing a corresponding Markov chain. Using this analysis, we relate the threshold-based strategies to Property 1 and Property 2, and construct an optimal solution to the problem (7) by choosing a specific threshold for the threshold-based strategies.

4.1 Threshold-Based Proactive Strategies

We describe the threshold-based strategies Ψ_P^ϕ in Algorithm 1. Specifically, we define $\phi \in (0, s]$ as the threshold parameter. When working proactively, the threshold-based strategy Ψ_P^ϕ works on request $J(t)$ at time t, where $J(t) = \min\{i \in \mathbf{\Pi}(t) | U_i(t) < \phi\}$ is the first request in the prediction window $\mathbf{\Pi}(t)$ which has not received ϕ bits of proactive service. By the definition of Ψ_P^ϕ, the process $\{J(t); t > 0\}$ is non-decreasing. In order to study the impact of ϕ on the threshold-based proactive strategies, we construct and analyze a corresponding Markov chain under given ϕ.

4.2 Markov Chain of System under Ψ_P^ϕ

We construct a Markov chain corresponding to the system under Ψ_P^ϕ, using methods applied in the analysis of M/G/1 queues and G/M/1 queues [20],

Algorithm 1 Threshold-based Strategies Ψ_P^{ϕ}

```
Main Procedure SYSTEM_RUN(φ)
    Choose the threshold as φ;
    Initialize V(t), Π(t)
    while t > 0 do
        if Request i arrives at t then
            Put reactive part S_i of request i into the tail of the queue V(t).
            Update prediction window Π(t)
        end if
        % Reactive work
        if V(t) > 0 then
            Transmit data from the head of the queue V(t) with full rate μ.
        end if
        % Proactive work
        if V(t) = 0 then
            Set J(t) = min{i ∈ Π(t) | U_i(t) < φ}
            % J(t) is the earliest potential request in Π(t) which has received less than φ bits of proactive service
            if U_{J(t)}(t) < φ then
                Transmit data of r_{J(t)} at full rate μ
            end if
        end if
    end while
End Procedure
```

Definition 3 (Markov Chain of the Proactive System under Ψ_P^{ϕ}). *Let* $T^{\phi} \triangleq (\tau_0, \tau_1, \tau_2, \ldots, \tau_n, \ldots)$ *be the sequence of transition epochs, where each* $\tau_n, n = 0, 1, \ldots$ *satisfies 1)* $V\left(\tau_n^+\right) = 0$*; 2)* $U_{J(\tau_n^+)}\left(\tau_n^+\right) = 0$*, and 3)* $J\left(\tau_n^+\right) > P\left(\tau_n^+\right)$*. The discrete-time process* $\{X_n : n = 0, 1, \ldots\}$ *with state space* $\{1, 2, 3, \ldots\}$ *is defined as:*

$$X_n = J\left(\tau_n^+\right) - P\left(\tau_n^+\right), n = 0, 1, 2, \ldots \tag{17}$$

In Proposition 2, we will show that $\{X_n\}$ is a Markov chain. We interpret the three conditions in Definition 3 as follows. Condition 1) means that there is no reactive traffic to serve right after τ_n, so the server can proactively serve requests in $\mathbf{\Pi}\left(\tau_n^+\right)$. Condition 2) means that at τ_n, the server starts to proactively work on request $J\left(\tau_n^+\right)$, which has not received proactive service before τ_n^+. The last condition means that the potential request to be proactively served at τ_n^+ should be a potential request which has not arrived in $\{P(t)\}$ by τ_n^+. To summarize, the discrete-time process $\{X_n : n = 0, 1, \ldots\}$ is constructed by sampling the system at $\{\tau_n : n = 0, 1, \ldots\}$ when the server starts to proactively work on a future potential request.

At each epoch $\tau_n, n \in \mathbb{Z}^+$, the nth *transition* in the Markov chain occurs. $X_n = J\left(\tau_n^+\right) - P\left(\tau_n^+\right), n = 0, 1, 2, \ldots$, represents how far the proactive service process $\{J(t); t \geq 0\}$ is ahead of the potential arrival process $\{P(t); t \geq 0\}$ at epoch τ_n^+. Figure 4 shows an example of how the transition epochs $\{\tau_n : n = 0, 1, \ldots\}$ are chosen.

Example: In the example shown in Figure 4, we choose $\phi = \frac{s}{2}$ in the threshold-based strategy. We make the following observations on the evolution of the process.

(1) No arrival occurs in (τ_0, τ_1). The server finishes proactively serving request 1 at τ_1, and starts to proactively serve request 2. The process in (τ_1, τ_2) evolves in the same way.
(2) In (τ_2, τ_3), the server proactively serves request 3; requests 1 and 2 arrive, with 1 realized and 2 not realized. At τ_3, the server starts to proactively serve request 4.
(3) In (τ_3, τ_4), request 3, 4, 5 arrive with only request 3 realized, before the server can finish proactively serving ϕ bits of request 4. Because the server cannot observe the arrival of request 4 or 5, it keeps proactively serving request 4 until τ'. At τ', the server starts to proactively serve request $5 \in \Pi(\tau') = (4, 5, \ldots)$. Nevertheless, condition 3) in Definition 3 is not satisfied at τ' $\left(J(\tau'^+) - P(\tau'^+) = 0 < 1\right)$. Thus, τ' is not a transition epoch. At τ_4, the

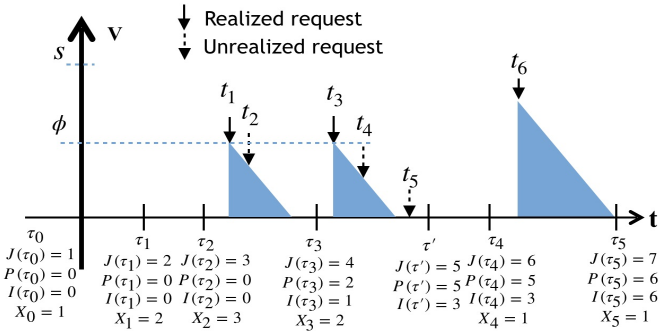

Fig. 4. Example: Transitions in the proactive system with Ψ_P^{ϕ}, with $\phi = \frac{s}{2}$

server starts to proactively serve request 6. Since conditions (1)-(3) in Definition 3 are all satisfied, τ_4 is a transition epoch.

(4) In (τ_4, τ_5), request 6 arrives and is realized before it receives ϕ bits proactively. Thus, it is served reactively until all bits are received. Since there is no arrival before it finishes, we have $I(\tau_5) = P(\tau_5) = 6$, so that the server starts proactively serving request 7, and τ_5 is a transition epoch.

We define $A_n \triangleq P\left(\tau_{n+1}^+\right) - P\left(\tau_n^+\right)$ as the number of potential arrivals in $(\tau_n, \tau_{n+1}]$, and $T_n \triangleq \tau_{n+1} - \tau_n$ as the nth inter-transition time. Starting from $X_n = x_n, x_n \in \mathbb{Z}^+$, the first $x_n - 1$ requests in $\mathbf{\Pi}(\tau_n^+)$, i.e., $P\left(\tau_n^+\right) + 1, \ldots, P\left(\tau_n^+\right) + x_n - 1$ (no requests if $x_n = 1$), have already received ϕ bits of proactive service by τ_n, and the request $P\left(\tau_n^+\right) + x_n$ just starts to be proactively served from τ_n^+. If $A_n \geq x_n$, we have $X_{n+1} = 1$. If $A_n < x_n$, X_{n+1} depends on A_n. In the following proposition, we formally describe the evolution of $\{X_n; n \geq 0\}$ and show its Markovian property.

Proposition 2. *The discrete-time process $\{X_n; n \geq 0\}$ defined in Definition 3 for the proactive system under Ψ_P^{ϕ} is Markovian, with the evolution*

$$X_{n+1} = \max\{X_n + 1 - A_n, 1\}, n = 0, 1, \ldots \tag{18}$$

Proof. Please refer to Appendix D for the proof. □

We now consider the transition probabilities $\{Pr\{X_{n+1} = x_{n+1} | X_n = x_n\} : x_n \in \mathbb{Z}^+, x_{n+1} \in \mathbb{Z}^+\}$.

(1) If $x_{n+1} > x_n + 1$, such transitions cannot happen by Definition 3.

(2) If $1 < x_{n+1} \leq x_n + 1$, we have the following fact:

$$Pr\{X_{n+1} = x_{n+1} | X_n = x_n\} = Pr\{A_n = x_n + 1 - x_{n+1} | X_n = x_n\}$$

which follows Proposition 2. If we let $A_n = k$, then $x_n > k \geq 0$.
An interesting fact is that:

$$Pr\{A_n = k | X_n = k + 1\} = Pr\{A_n = k | X_n = k + 2\} = \ldots \tag{19}$$

This is because τ_{n+1} is determined by 1) arrival epochs $\left\{t_i : i > P(\tau_n^+), i \in \mathbb{Z}^+\right\}$, 2) realizations $\left\{R_i : i > P(\tau_n^+), i \in \mathbb{Z}^+\right\}$, and 3) reactive work to be done for each request $\left\{S_i : i > P(\tau_n^+), i \in \mathbb{Z}^+\right\}$.

If $A_n < x_n$, we have:

$$\begin{aligned}\text{If } x_n > 1 : & U_i(\tau_n^+) = \phi, i = P(\tau_n^+) + 1, P(\tau_n^+) + 2, \ldots, P(\tau_n^+) + x_n - 1 \\ & U_i(\tau_n^+) = 0, i = P(\tau_n^+) + x_n, P(\tau_n^+) + x_n + 1, \ldots \\ \text{If } x_n = 1 : & U_i(\tau_n^+) = 0, i = P(\tau_n^+) + 1, P(\tau_n^+) + 2, \ldots\end{aligned}$$

by the definition of threshold-based strategies. Then for these $A_n = k < x_n$ arrivals, we have:

$$S_i = s - \phi, i = P(\tau_n^+) + 1, P(\tau_n^+) + 2, \ldots, P(\tau_n^+) + k, \text{ if } k = 1, 2, \ldots$$

And there is no reactive work to be done if there is no arrival before next transition, i.e., $A_n = 0$. So $Pr\{A_n = k | X_n = x_n\}$ only depends on conditions 1) and 2) if $k < x_n$, which implies (19).
Here we define:

$$p_k^\phi \triangleq Pr\{A_n = k | X_n = x_n\}, \forall x_n > k \tag{20}$$

and we have

$$Pr\{X_{n+1} = x_{n+1} | X_n = x_n\} = p^\phi_{x_n+1-x_{n+1}}, \text{if } 1 < x_{n+1} \leq x_n + 1 \tag{21}$$

(3) If $x_{n+1} = 1$, we have:

$$\begin{aligned}Pr\{X_{n+1} = x_{n+1} | X_n = x_n\} &= 1 - \sum_{i=2}^{\infty} Pr\{X_{n+1} = i | X_n = x_n\} \\ &= 1 - \sum_{i=2}^{x_n+1} p^\phi_{x_n+1-i} \\ &= 1 - \sum_{k=0}^{x_n-1} p_k^\phi \\ &= \sum_{k=x_n}^{\infty} p_k^\phi\end{aligned} \tag{22}$$

Then the transition probabilities can be written as:

$$\begin{aligned}p^\phi_{x_n x_{n+1}} &\triangleq Pr\{X_{n+1} = x_{n+1} | X_n = x_n\} \\ &= \begin{cases} 0, & x_{n+1} \geq x_n + 2 \\ p^\phi_{x_n+1-x_{n+1}}, & 1 < x_{n+1} \leq x_n + 1 \\ \sum_{k=x_n}^{\infty} p_k^\phi, & x_{n+1} = 1 \end{cases} \\ &, \forall x_n \in \mathbb{Z}^+, \forall x_{n+1} \in \mathbb{Z}^+\end{aligned} \tag{23}$$

Or equivalently, we can write the transition probabilities in matrix form:

$$P^\phi = \begin{bmatrix} \sum_{k=1}^{\infty} p_k^\phi & p_0^\phi & & \\ \sum_{k=2}^{\infty} p_k^\phi & p_1^\phi & p_0^\phi & \\ \sum_{k=3}^{\infty} p_k^\phi & p_2^\phi & p_1^\phi & p_0^\phi \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{24}$$

where the empty entries are 0. Notice that it is structurally similar to the transition probability matrix of the Markov chain of G/M/1 queue in [20].

Although we have developed the structure of the transition probability matrix, the probabilities $\left\{p_k^{\phi} : k = 0, 1, 2, \ldots\right\}$ are still unknown. In the following theorem, we are going to prove an important result of the probabilities.

Proposition 3. *The probabilities* $\left\{p_k^{\phi} : k = 0, 1, 2, \ldots\right\}$ *satisfy the following relationships:*

$$\sum_{k=0}^{\infty} p_k^{\phi}(1-k) \begin{cases} > 0, if & \phi < U^* \\ = 0, if & \phi = U^* \\ < 0, if & \phi > U^* \end{cases} \tag{25}$$

where $U^* = \frac{\mu - p\lambda s}{\lambda(1-p)}$, *as defined in (15).*

Proof. Please refer to Appendix E for the proof. □

Although we obtain some knowledge about transition probabilities of the Markov chain from Proposition 3, a remaining problem of the Markov Chain is the distribution of T_n and A_n. If $T_n = \infty$ with a positive probability, the next transition may never happen. Therefore, we have the following proposition on the expectations of T_n and A_n.

Proposition 4. *In the Markov Chain of the proactive system with* Ψ_P^{ϕ} *as defined in Definition 3, we have:*

$$E[T_n | X_n = x_n] < \infty,\ E[A_n | X_n = x_n] < \infty,\ \forall x_n \in \mathbb{Z}^+,\ \forall \phi \in (0, s] \tag{26}$$

Proof. Please refer to Appendix F for the proof. □

Proposition 4 implies that $Pr\{T_n < \infty\} = 1, Pr\{A_n < \infty\} = 1, \forall n \in \mathbb{Z}^+$. Therefore transitions in the corresponding Markov chain will almost surely happen in finite time.

To investigate the asymptotic behavior of the system, we need to characterize the recurrence of the Markov chain of the system. Based on Proposition 3 and Proposition 4, we have the following theorem on the recurrence of the Markov chain of the proactive system under Ψ_P^{ϕ}.

Theorem 2. *The Markov chain of the proactive system with* Ψ_P^{ϕ} *is 1) transient if* $\phi < U^*$, *2) positive recurrent if* $\phi > U^*$, *and 3) null recurrent if* $\phi = U^*$.

Proof. From Proposition 3, we can easily prove that:

$$\sum_{k=0}^{\infty} p_k^{\phi} k \begin{cases} < 1, if & \phi < U^* \\ = 1, if & \phi = U^* \\ > 1, if & \phi > U^* \end{cases} \tag{27}$$

In Section 10.3.3 of [14], the relation between $\sum_{k=0}^{\infty} p_k^{\phi} k$ and the recurrence of the corresponding Markov chain is discussed. To be specific, the conclusion is that the Markov chain is 1) positive recurrent if $\sum_{k=0}^{\infty} p_k^{\phi} k > 1$, 2) null recurrent if $\sum_{k=0}^{\infty} p_k^{\phi} k = 1$, and 3) transient if $\sum_{k=0}^{\infty} p_k^{\phi} k < 1$. Our conclusion directly follows. □

Theorem 2 characterizes the relationship between ϕ and the recurrence of the Markov chain under Ψ_P^{ϕ}. The recurrence of the corresponding Markov chain under different ϕ is the crucial key to investigating the relationship between Property 1 and Property 2 with the threshold-based strategies. In the following, we are going to discuss this relationship.

Property 1 of the Threshold-based Strategies: First, we focus on Property 1 and the threshold-based strategies in the following lemma.

Lemma 1. *A threshold-based strategy satisfies Property 1 if and only if the corresponding Markov chain satisfies:*

$$\lim_{n\to\infty} \frac{X_n}{n} = 0, \ w.p.1$$

Proof. Please refer to Appendix G for the proof. □

Lemma 1 transforms the conditions for Property 1 from the continuous sense in Definition 1 to a discrete condition based on transitions in the Markov chain. The term $\lim_{n\to\infty} \frac{X_n}{n}$ is closely related to the recurrence of the Markov chain, which has been characterized in Theorem 2. Then we have the following theorem on Property 1 of the threshold-based strategies.

Theorem 3. *A threshold-based strategy satisfies Property 1 if and only if $\phi \geq U^*$.*

Proof. Please refer to Appendix H for the proof. □

Another way of stating Theorem 3 is that a threshold-based strategy satisfies Property 1, if and only if the corresponding Markov chain is recurrent. Recall that the states X_n's represent the gaps between the proactive service process $\{J(t); t > 0\}$ and the potential process $\{P(t); t > 0\}$. If the corresponding Markov chain is recurrent, the state $X_n = 1$ will always happen. This implies that the proactive service done effectively reduces the reactive traffic of the requests which have arrived, which is also the insights of Property 1.

Property 2 of the Threshold-based Strategies: Next, we are going to discuss Property 2 of the threshold-based strategies. As we discussed in Proposition 1, $\overline{U} \geq \overline{U_A}$ is true due to our service model. Predictions are likely to receive more proactive service if they are unrealized. Because of our assumptions on the orderliness of predictions in $\mathbf{\Pi}(t)$, the predictions which have arrived but not realized are always the earliest predictions in $\mathbf{\Pi}(t)$. Intuitively, a threshold-based strategy with a larger ϕ, which prefers to serve the earliest predictions in $\mathbf{\Pi}(t)$, is more likely to achieve $\overline{U} > \overline{U_A}$. We rigorously characterize the relationship of the threshold-based strategies and Property 2 in the following theorem.

Theorem 4. *The threshold-based strategy Ψ_P^{ϕ} satisfies Property 2 if and only if $\phi \leq U^*$.*

Proof. Please refer to Appendix I for the proof. □

Theorem 4 verified our previous intuitions. Similar to Theorem 3, Theorem 4 has an equivalent statement: the threshold-based strategy Ψ_P^{ϕ} satisfies Property 2, if and only if the corresponding Markov chain is *NOT* positive recurrent. As we discussed, $\overline{U} > \overline{U_A}$ is more likely to happen when the strategy proactively works on the requests which have arrived but not realized. This only happens when the system transits to state $X_n = 1$. In a transient or null recurrent case, the system state $X_n = 1$ does not happen comparably often as n. As a result, Property 2 is satisfied in these cases.

Based on Property 1 and Property 2 of the threshold-based strategies as characterized in Theorem 3 and 4, we have the following corollary which solves the optimization problem (7).

Corollary 2. *$\overline{U}$ in (7) is maximized with a threshold-based proactive strategy Ψ_P^{ϕ} if and only if $\phi = U^*$.*

Proof. By combining Theorem 3, Theorem 4 and Corollary 1, the corollary directly follows. □

Based on the corollary, $\Psi_P^{U^*}$ is a solution to the optimization problem (7). Notice that this is the only threshold-based strategy which maximizes $\overline{U}$, and it is the only case where the corresponding Markov chain is null recurrent.

We obtained the following valuable insights about the characteristics of an optimal proactive strategy under prediction uncertainties. First, the strategy should not overemphasize predictions which are near in the future, as how the EDF strategy works, in order to account for the fact that the potential requests may not be realized. Second, it should not overemphasize predictions which are far in the future, in order to provide sufficient proactive services for the requests which may arrive in the near future. Balancing these two effects as a function of the prediction uncertainties is the key to designing a desirable proactive strategy.

5 DELAY COMPARISON BETWEEN UNIFORM AND EDF STRATEGIES

In this section, we focus on two special proactive strategies, which are the EDF (Earliest-Deadline-First) type strategy and the UNIFORM strategy. The EDF strategy can be seen as the threshold-based strategy with $\phi = s$, which means the server will always first proactively work on the first request in $\mathbf{\Pi}(t)$ which has not been completely proactively served. The EDF strategy has been widely used in many scheduling problems in queueing systems. Intuitively, reducing traffic at the beginning of a congested period might be the most efficient way to reduce delay. In our case where all objects have a uniform size, the EDF strategy works the same as the shortest remaining time first (SRTF) strategy, which achieves the optimal delay in a reactive queueing system. In a proactive system, the authors of [9] have proved that the EDF strategy can achieve asymptotic optimality in terms of delay when the size of the prediction window goes to infinity with full knowledge of future requests and their arrival epochs. However, we will show that the UNIFORM strategy outperforms the EDF strategy in terms of delay in the case with uncertain predictions.

First, we derive an important property of the UNIFORM strategy in the following corollary.

Corollary 3. *Given μ, λ, s and p as system parameters which satisfy $\frac{\mu}{s} \leq \lambda < \frac{\mu}{ps}$, the system operates under the UNIFORM strategy $\Psi_P^{U^*}$. Then the limiting empirical distribution of U_i satisfies*

$$\lim_{t\to\infty} \frac{1}{I(t)} \sum_{i=1}^{I(t)} \mathbb{1}\left(U_i = U^*\right) = 1, w.p.1 \tag{28}$$

Proof. Please refer to Appendix J for the proof. □

The Corollary 3 shows that the requests under the UNIFORM strategy receive U^* bits of proactive service with probability 1. Consequently, the reactive work of each actual request is S^* with probability 1. Since almost all actual requests receive the same amount of proactive service, we call this strategy UNIFORM. In the following, we derive the closed-form expression for the average delay per actual request under the UNIFORM strategy.

Corollary 4. *Given μ, λ, s and p as system parameters which satisfy $\frac{\mu}{s} \leq \lambda < \frac{\mu}{ps}$, the average delay $\overline{D}_{U(NIFORM)}$ per actual request under the UNIFORM strategy $\Psi_P^{U^*}$ can be expressed as:*

$$\overline{D}_U = \frac{(\lambda s - \mu)(2\mu - \mu p - \lambda p s)}{2\mu\lambda(1-p)(\mu - \lambda p s)}, w.p.1 \tag{29}$$

If we define $\overline{D}_{R(eactive)}$ as the average delay of each actual request under the reactive scheme, the ratio of $\frac{\overline{D}_U}{\overline{D}_R}$ can be expressed as:

$$\frac{\overline{D}_U}{\overline{D}_R} = \frac{(\lambda s - \mu)(2\mu - \mu p - \lambda p s)}{\lambda s(1-p)(2\mu - s\lambda p)}, w.p.1 \tag{30}$$

Proof. Please refer to Appendix K for the proof. □

The ratio in (30) directly compares the delay of UNIFORM strategy against the reactive scheme, and we will plot it in Section 6. Next, we compare the average delay of the UNIFORM strategy against EDF strategy.

Corollary 5. *Given μ, λ, s and p as system parameters which satisfy $\frac{\mu}{s} \leq \lambda < \frac{\mu}{ps}$, the average delay of UNIFORM strategy $\overline{D}_{U(NIFORM)}$ is no greater than the EDF strategy $\overline{D}_{E(DF)}$ with probability 1:*

$$\overline{D}_U \leq \overline{D}_E, w.p.1 \tag{31}$$

The equality holds if and only if $p = 0$.

Notice that $0 \leq p < \frac{\mu}{\lambda s} < 1$, so $p = 0$ is the only value where the equality holds.

Proof. Please refer to Appendix L for the proof. □

The proof of Corollary 5 reveals the insights on why the UNIFORM strategy outperforms the EDF strategy. First, the EDF strategy satisfies Property 1 but not Property 2. As a result, the average reactive work per actual request $\overline{S}$ is larger under the EDF strategy by Corollary 2, which means the server needs to deal with more reactive work on average. Second, the unbalanced allocation of proactive rates in the EDF strategy impacts the delay performance. As shown in Figure 3, the EDF strategy works well when requests are realized, like the first 5 requests. However, when the first future potential request seen by the server is not realized, the EDF strategy usually achieves awful delay performance. Also take Figure 3 as an example. Request 6 receives a lot of proactive services but it is not realized, which causes request 7 to be served almost completely reactively. Consequently, request 8 suffers from large queueing delay.

6 NUMERICAL EVALUATION

We perform extensive experiments to study the delay performance of threshold-based strategies. Specifically, we compare the UNIFORM strategy with the EDF strategy, with the reactive scheme as a baseline. In our simulations, we consider the network in Figure 1. We set $\mu = 10$ and $s = 1$ in all of our experiments.

In our simulations, we gradually increase the threshold ϕ from 0 to s and compare the average delay per actual request in each case. Specifically, when $\phi = s$, the strategy becomes the EDF strategy; when $\phi = U^*$, the strategy becomes the UNIFORM strategy; and when $\phi = 0$, the system operates in the reactive scheme. The term λp determines how heavily the network is loaded, and we choose $\lambda p = 6$ as the lightly-loaded network scenario and $\lambda p = 9.6$ as the heavily-loaded network scenario. With each fixed value of λp, we gradually increase λ from 10 to 20 and choose p correspondingly, to evaluate the effects of prediction uncertainties on the delay performance. We set the simulation time to be 10^7 seconds.

6.1 Infinite Prediction Window Scenarios

We first demonstrate the delay performance of threshold-based strategies under an infinite prediction window.

Figures 5 and 6 show the delay performance of threshold-based strategies, with different thresholds and different prediction uncertainties. The x-axis represents the threshold ϕ, which gradually increases from 0 to s. Each curve corresponds to a (λ, p) combination with the same product λp. Each vertical dotted line represents the thresholds U^* of the UNIFORM strategy under each (λ, p) combination, which shares same color with the corresponding curve. For each curve, the delay of the EDF strategy is shown at $x = s = 1$, and the delay of the reactive scheme is shown at $x = 0$.

Here are some interesting observations on the plots:

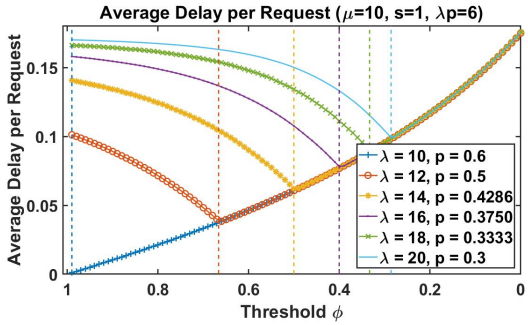

Fig. 5. Comparisons among threshold-based methods: $\lambda p = 6$

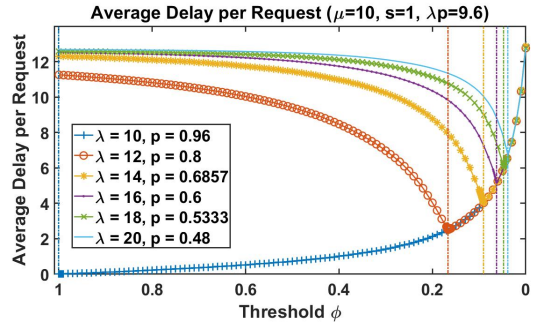

Fig. 6. Comparisons among threshold-based methods: $\lambda p = 9.6$

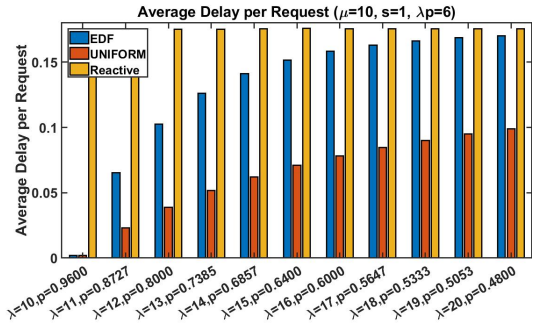

Fig. 7. Comparisons among EDF,UNIFORM and Reactive Schemes: $\lambda p = 6$

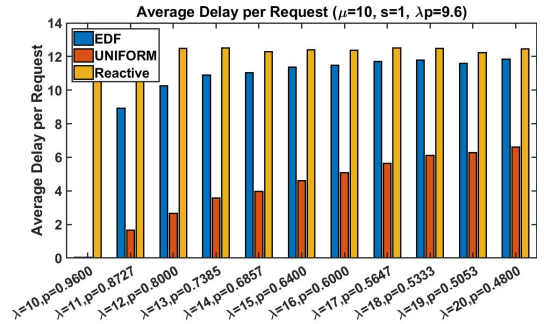

Fig. 8. Comparisons among EDF,UNIFORM and Reactive Schemes: $\lambda p = 9.6$

- The vertical lines perfectly mark the minimum point on each curve.[3] This implies that the UNIFORM strategy always achieves the best delay performance among all the threshold-based strategies.
- If we compare two curves corresponding to different (λ, p) combinations (but with the same product λp), we can see that the delay performance of the curve with larger p and smaller λ outperforms the one with smaller p and larger λ, until they overlap. This is because larger p and smaller λ imply higher predictability, so that the proactive strategy has the potential to achieve a more desirable delay performance. The overlapping part is due to the choice of an overly-small threshold ϕ. In this case, almost every request receives ϕ bits of proactive service, even in the case with higher predictability. This points to the significance of Property 1.
- If we compare Figure 5 and Figure 6, we observe that the curves between $\phi = s = 1$ and $\phi = U^*$ are flatter in the lightly-loaded scenario ($\lambda p = 6$). This implies that delay performance is less sensitive to threshold ϕ when the network is less congested. In the heavily-loaded case, the choice of threshold ϕ is more crucial for achieving desirable delay performance.

In order to make more straightforward comparisons among the EDF strategy, the UNIFORM strategy and the reactive scheme, we plot the average delay achieved by these strategies in Figures

[3]Note that the vertical lines indicate $\phi = 1$ on the curves for $\lambda = 10$ in both figures, meaning that $\phi = 1$ is the optimal threshold, i.e. the UNIFORM strategy is the same as the EDF strategy in this case.

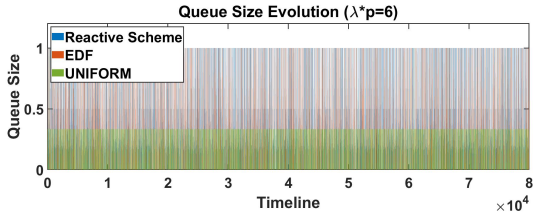

Fig. 9. Queue Size Evolution Comparisons among Reactive Scheme, EDF strategy and UNIFORM strategy: $\lambda p = 6$

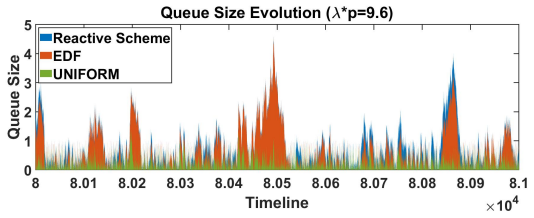

Fig. 10. Queue Size Evolution Comparisons among Reactive Scheme, EDF strategy and UNIFORM strategy: $\lambda p = 9.6$

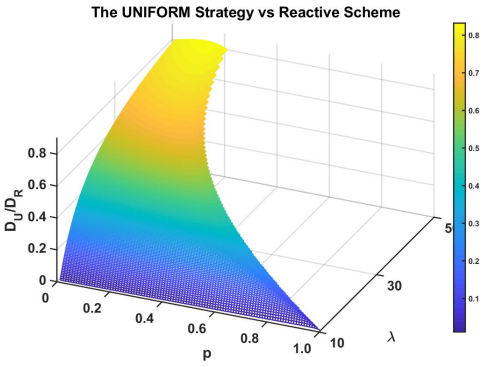

Fig. 11. Theoretical Delay Comparison between UNIFORM Strategy and Reactive Scheme

7 and 8. We observe that with the delay performance of the reactive scheme as the baseline, the delay performance of the EDF strategy becomes much worse in the heavily-loaded scenario as compared with that in the light-loaded scenario, whereas the delay performance of the UNIFORM strategy is relatively stable in both scenarios.

In Figures 9 and 10, we compare queue size evolutions under the EDF strategy, the UNIFORM strategy and reactive scheme. In Figure 9, we observe that the queue size under the EDF strategy is very similar to that under the reactive scheme, implying that in this case, many requests do not receive proactive service under the EDF strategy. On the other hand, the UNIFORM strategy is able to keep the queue size at a low level. This is because the EDF strategy assigns proactive service in a very unbalanced manner, while the UNIFORM strategy assigns proactive resources almost uniformly among all requests. In Figure 10, the differences are magnified. When the network is heavily loaded, the EDF strategy fails to effectively control congestion, but the UNIFORM strategy is able to steadily keep the queue size at a very low level. This difference directly leads to the gap between the delay performance of the EDF strategy and the UNIFORM strategy in the heavily-loaded scenario.

In Figure 11, we plot the ratio of the average delay under the UNIFORM strategy to that of the reactive scheme, as calculated in Corollary 4. In this plot, λ is chosen to be from 10 to 50, and p is chosen from 0 to $10/\lambda$ (so that $\lambda ps < \mu$). For a fixed λ, the system is more congested with a larger p. As can be observed, the UNIFORM strategy achieves a consistent advantage over the reactive scheme with a fixed λ. Even in a very congested case with bad predictions ($\lambda = 50$ and p approaches 0.2), the UNIFORM strategy still can achieve approximately a 20% advantage over the reactive scheme.

6.2 Finite Prediction Window Scenarios

In practice, prediction algorithms can only predict user requests in a finite future. In this section, we experimentally study the impact of prediction window size on delay performance. A finite prediction

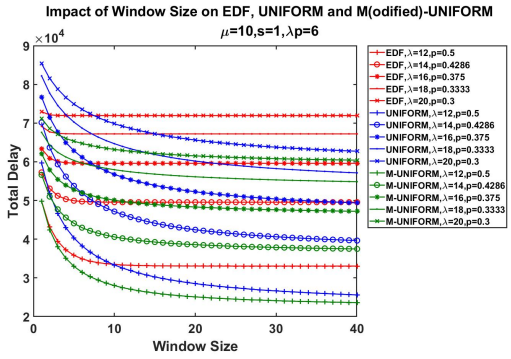

Fig. 12. Comparisons among EDF,UNIFORM and Modified-UNIFORM: $\lambda p = 6$

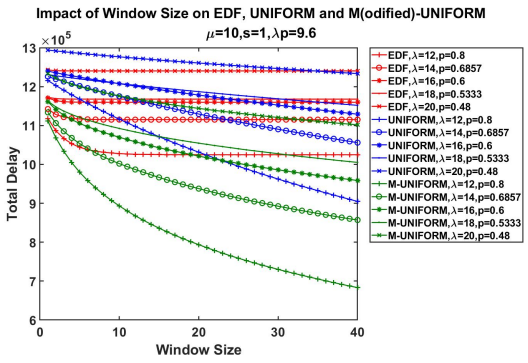

Fig. 13. Comparisons among EDF,UNIFORM and Modified-UNIFORM: $\lambda p = 9.6$

window $\mathbf{\Pi}(t) = (I(t)+1, I(t)+2, \ldots, I(t)+W)$ is considered, where only W predictions are available for any $t > 0$. In this case, there is a possibility that all the potential requests in $\mathbf{\Pi}(t)$ have been proactively served with ϕ bits. When this happens, the system must remain idle until there are new predictions available.

We carried out a series of experiments to assess the impact of the prediction window size W on the delay performance of EDF and UNIFORM strategies. We also consider a Modified-UNIFORM (M-UNIFORM) strategy, as described in Algorithm 2. After every available prediction in $\mathbf{\Pi}(t)$ receives ϕ bits of proactive service, the M-UNIFORM strategy starts to proactively serve the earliest request which has not received s bits of proactive service in the prediction window.

Figures 12 and 13 show the delay performance of these strategies. The delay performance of the EDF strategy converges faster with respect to W. Thus the EDF strategy does not require a large prediction window to achieve its best delay performance. On the other hand, the UNIFORM strategy converges much more slowly, especially in the heavily-loaded case. It also requires a moderately large prediction window size for the UNIFORM strategy to outperform the EDF strategy, especially in the heavily-loaded case. However, we can greatly improve the delay performance of the UNIFORM strategy with a few simple modifications. We can observe that the performance of the M-UNIFORM strategy in the small-window region is greatly improved over that of the UNIFORM strategy. For instance, in Figure 13, the UNIFORM strategy requires a window size W greater than 34 to outperform the EDF strategy for the case of $\lambda = 12, p = 0.8$. However, the M-UNIFORM strategy outperforms the EDF strategy even when $W = 1$.

7 CONCLUSIONS

In this paper, we looked into the fundamental queueing dynamics of proactive caching strategies under uncertain predictions and developed insights on how to design a proactive strategy to achieve desirable delay performance in a single queue system. We solved an optimization problem of maximizing the limiting average amount of proactive service per request. By comparing queueing dynamics in the proactive scheme and reactive scheme under the same sample path, we derived a tight upper bound on the objective with uncertain predictive information of future requests. We proposed a family of threshold-based strategies, and constructed the Markov chain of the system to analyze the asymptotic behavior of the proactive system. Consequently, we found the optimal strategy, i.e. the UNIFORM strategy, by properly choosing the threshold in the threshold-based strategies, which corresponds to a null recurrent Markov chain. We obtained important insights

Algorithm 2 Modified UNIFORM Strategy

Main Procedure SYSTEM_RUN(U^*)
  Choose the threshold as U^*;
  Initialize $V(t)$, $I(t)$, $\mathbf{\Pi}(t)$
  while $t > 0$ **do**
    if Request i arrives at t **then**
      Put reactive part S_i of request i into the tail of the queue $V(t)$.
      Update prediction window $\mathbf{\Pi}(t)$
    end if
    if $V(t) > 0$ **then**
      % Reactive work
      Transmit data from the head of the queue $V(t)$ with full rate μ.
    end if
    if $V(t) = 0$ **then**
      % Proactive work
      Set $i = \min\{i | I(t) < i \leq I(t) + W, U_i(t) < U^*\}$
      % i is the earliest potential request in $\mathbf{\Pi}(t)$ which has received less than U^* bits of proactive service
      if $i == null$ **then**
        %All potential requests in $\mathbf{\Pi}(t)$ have received U^* bits of proactive work
        Set $j = \min\{i | I(t) < i \leq I(t) + W, U_j(t) < s\}$
        if $U_j(t) < s$ **then**
          Transmit data of r_j with full rate μ
        end if
        if $j == null$ **then**
          %Every request in $\mathbf{\Pi}(t)$ has received s units of proactive work
          Stay idle
        end if
      else
        if $U_i(t) < U^*$ **then**
          Transmit data of r_i with full rate μ
        end if
      end if
    end if
  end while
End Procedure

about the characteristics of an optimal proactive strategy: the strategy should balance the amount of proactive work between the potential requests which are arriving sooner and the ones arriving later, based on the uncertainties in predictions. We derived the closed-form expression of average delay per actual request under the UNIFORM strategy, and analytically compared it with the commonly used EDF type strategy. We showed that the UNIFORM strategy outperforms the EDF strategy in all the non-trivial scenarios, which is verified by extensive numerical experiments under differently congested network scenarios. Experimental results also showed that delay can be dramatically decreased by proactive caching techniques not only in the lightly-loaded region as claimed in [22], but also in the heavily-loaded case if properly designed. Our work provides valuable insights on how to optimally design a proactive strategy to improve the delay performance in the system.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from National Science Foundation grants CNS-1718355, OAC-1659403, CNS-NeTS-1514260, CNS-NeTS-1717045, CMMI-SMOR-1562065, CNS-ICN-WEN-1719371, and CNS-SpecEES-1824337, and from DTRA grant HDTRA1-14-1-0058, as well as from research grants by Intel Corp. and Cisco Systems.

REFERENCES

[1] Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. 2013. A Peek into the Future: Predicting the Evolution of Popularity in User Generated Content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 607–616. https://doi.org/10.1145/2433396.2433473

[2] Faisal Alotaibi, Sameh Hosny, John Tadrous, Hesham El Gamal, and Atilla Eryilmaz. 2015. Towards A Marketplace for Mobile Content: Dynamic Pricing and Proactive Caching. arXiv:1511.07573[cs.GT].

[3] Matthew Andrews. 2000. Probabilistic end-to-end delay bounds for earliest deadline first scheduling. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, Vol. 2. IEEE, 603–612.
[4] E. Bastug, M. Bennis, and M. Debbah. 2014. Living on the edge: The role of proactive caching in 5G wireless networks. *IEEE Communications Magazine* 52, 8 (Aug 2014), 82–89. https://doi.org/10.1109/MCOM.2014.6871674
[5] Dimitri P Bertsekas, Robert G Gallager, and Pierre Humblet. 1992. *Data networks*. Vol. 2. Prentice-Hall International New Jersey.
[6] Kun Chen and Longbo Huang. 2018. Timely-throughput optimal scheduling with prediction. *IEEE/ACM Transactions on Networking* (2018).
[7] Cisco. 2017. The Zettabyte Era: Trends and Analysis. *White Paper* (2017).
[8] Leonidas Georgiadis, Michael J. Neely, and Leandros Tassiulas. 2006. Resource Allocation and Cross-Layer Control in Wireless Networks. *Foundations and Trends in Networking* 1, 1 (2006), 1–144. https://doi.org/10.1561/1300000001
[9] Longbo Huang, Shaoquan Zhang, Minghua Chen, Xin Liu, Longbo Huang, Shaoquan Zhang, Minghua Chen, and Xin Liu. 2016. When Backpressure Meets Predictive Scheduling. *IEEE/ACM Trans. Netw.* 24, 4 (Aug. 2016), 2237–2250. https://doi.org/10.1109/TNET.2015.2460749
[10] Stratis Ioannidis and Edmund Yeh. 2018. Adaptive Caching Networks With Optimality Guarantees. *IEEE/ACM Trans. Netw.* 26, 2 (April 2018), 737–750. https://doi.org/10.1109/TNET.2018.2793581
[11] Mehdi Kargahi and Ali Movaghar. 2006. A method for performance analysis of earliest-deadline-first scheduling policy. *The Journal of Supercomputing* 37, 2 (2006), 197–222.
[12] Ron Kohavi and Roger Longbotham. 2007. Online Experiments: Lessons Learned. *Computer* 40, 9 (Sept 2007), 103–105. https://doi.org/10.1109/MC.2007.328
[13] Milad Mahdian and Edmund Yeh. 2017. MinDelay: Low-latency Forwarding and Caching Algorithms for Information-Centric Networks. arXiv:1710.05130[cs.NI].
[14] Sean P. Meyn and Richard L. Tweedie. 1993. Markov chains and stochastic stability.
[15] Leela Srikar Muppirisetty, John Tadrous, Atilla Eryilmaz, and Henk Wymeersch. 2015. On proactive caching with demand and channel uncertainties. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 1174–1181. https://doi.org/10.1109/ALLERTON.2015.7447141
[16] Henrique Pinto, Jussara M. Almeida, and Marcos A. Gonçalves. 2013. Using Early View Patterns to Predict the Popularity of Youtube Videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 365–374. https://doi.org/10.1145/2433396.2433443
[17] Vijay Sivaraman and Fabio Chiussi. 2000. Providing end-to-end statistical delay guarantees with earliest deadline first scheduling and per-hop traffic shaping. In *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE, 631–640.
[18] John Tadrous and Atilla Eryilmaz. 2016. On Optimal Proactive Caching for Mobile Networks With Demand Uncertainties. *IEEE/ACM Transactions on Networking* 24, 5 (October 2016), 2715–2727. https://doi.org/10.1109/TNET.2015.2478476
[19] John Tadrous, Atilla Eryilmaz, and Hesham El Gamal. 2013. Proactive resource allocation: Harnessing the diversity and multicast gains. *IEEE Transactions on Information Theory* 59, 8 (2013), 4833–4854.
[20] Ronald W Wolff. 1989. *Stochastic modeling and the theory of queues*. Pearson College Division.
[21] Edmund Yeh, Tracey Ho, Ying Cui, Michael Burd, Ran Liu, and Derek Leong. 2014. VIP: A Framework for Joint Dynamic Forwarding and Caching in Named Data Networks. In *Proceedings of the 1st ACM Conference on Information-Centric Networking (ACM-ICN '14)*. ACM, New York, NY, USA, 117–126. https://doi.org/10.1145/2660129.2660151
[22] Shaoquan Zhang, Longbo Huang, Minghua Chen, and Xin Liu. 2017. Proactive Serving Decreases User Delay Exponentially: The Light-Tailed Service Time Case. *IEEE/ACM Trans. Netw.* 25, 2 (April 2017), 708–723. https://doi.org/10.1109/TNET.2016.2607840

A PROOF OF PROPOSITION 1

First we consider the terms $\frac{\sum_{i=1}^{I(t)} U_i(t_i)}{I(t)}$ and $\frac{\sum_{i\in\mathbb{Z}^+:R_i=1,i\le I(t)} U_i(t_i)}{A(t)}$. $\frac{\sum_{i=1}^{I(t)} U_i(t_i)}{I(t)}$ is the average of terms in $\{U_i(t_i): i \le I(t)\}$. $\frac{\sum_{i\in\mathbb{Z}^+:R_i=1,i\le I(t)} U_i(t_i)}{A(t)}$ is the average of the samples in $\{U_i(t_i): i \le I(t), R_i = 1\}$, which are selected from $\{U_i(t_i): i \le I(t)\}$ if $R_i = 1$. One important fact is that $U_i(t_i)$ is independent of R_i, because the server has no knowledge of R_i before t_i. Because R_i's are IID, we have:

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i(t_i)}{I(t)} = \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1,i\le I(t)} U_i(t_i)}{A(t)}, w.p.1 \tag{32}$$

Recall that $U_i = U_i(t_i)$ if $R_i = 1$, and $U_i \geq U_i(t_i)$ if $R_i = 0$. So we have:

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i}{I(t)} \geq \lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i(t_i)}{I(t)} \tag{33}$$

$$\lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\leq I(t)} U_i}{A(t)} = \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\leq I(t)} U_i(t_i)}{A(t)} \tag{34}$$

By combining the equations above, we have:

$$\lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i}{I(t)} \geq \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\leq I(t)} U_i}{A(t)}, w.p.1 \tag{35}$$

Therefore by definitions of $\overline{U}$ and $\overline{U_A}$, we have $\overline{U} \geq \overline{U_A}$, $w.p.1$.

B PROOF OF THEOREM 1

First we make the following definitions similar to (9), (10) and (11):

The amount of time that Ψ_R works in idle state (namely Reactive Idle) from 0 to t is:

$$T_{RI}(t) \triangleq |\{\tau \in (0,t] : V(\tau) = 0\}| \tag{36}$$

The amount of time that Ψ_R works in busy state (namely Reactive Busy) from 0 to t is:

$$T_{RB}(t) \triangleq |\{\tau \in (0,t] : V(\tau) > 0\}| \tag{37}$$

The limiting fraction of time that Ψ_R works in idle state is:

$$\alpha_{RI} \triangleq \lim_{t\to\infty} \frac{T_{RI}(t)}{t} \tag{38}$$

The limiting fraction of time that Ψ_R works in busy state is:

$$\alpha_{RB} \triangleq \lim_{t\to\infty} \frac{T_{RB}(t)}{t} \tag{39}$$

Compare reactive scheme with proactive scheme under the same sample path. Define the system state at time t as "*Proactive Served*", if Ψ_R works in busy state at time t, and Ψ_P works in proactive state at time t. The amount of time that Ψ_P works in "Proactive Served" state is:

$$T_{PS}(t) \triangleq |\{\tau \in (0,t] : V_P(\tau) = 0, V_R(\tau) > 0\}| \tag{40}$$

where $V_P(t)$ is the unfinished work in proactive scheme at t and $V_R(t)$ is the unfinished work in reactive scheme at t. The corresponding time intervals are marked in Figure 3.

Observe the system at time t when $V(t) = 0$ in both reactive scheme and proactive scheme. All the potential requests in $\{i : t_i < t\}$, the corresponding realizations $\{R_i : t_i < t\}$ and the resulting $\{U_i(t) : t_i < t\}$ of strategy Ψ_P have been determined, so the entire timeline from 0 to t can be divided in two states in both reactive scheme and proactive scheme, as shown in Figure 3. Consequently, we have:

$$T_{RB}(t) + T_{RI}(t) = T_{PR}(t) + T_{PP}(t) = t \tag{41}$$

An important fact to be noticed here is that:

$$T_{PP}(t) = T_{PS}(t) + T_{RI}(t) \tag{42}$$

Then by (40):

$$\mu T_{PS}(t) = \sum_{i\in\mathbb{Z}^+:R_i=1, i\leq I(t)} U_i(t) \tag{43}$$

where the term $\sum_{i\in\mathbb{Z}^+:R_i=1,\, i\le I(t)} U_i(t)$ is the total amount of proactive work received by all the actual requests arrived in $(0,t)$.

Next, the total proactive work done by time t equals $\mu T_{PP}(t)$ by the definition of $T_{PP}(t)$, which satisfies the follow equation:

$$\mu T_{PP}(t) = \sum_{i=1}^{I(t)} U_i(t) + \sum_{i=I(t)+1}^{\infty} U_i(t) \tag{44}$$

where the term $\sum_{i=1}^{I(t)} U_i(t)$ is the total proactive work done for requests in $\{i\in\mathbb{Z}^+ : i\le I(t)\}$, and $\sum_{i=I(t)+1}^{\infty} U_i(t)$ is the total proactive work done for requests in $\{i\in\mathbb{Z}^+ : i > I(t)\}$.

Next we have:

$$\begin{aligned}
\frac{\lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1,\, i\le I(t)} U_i(t)}{t}}{\lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i(t)}{t}} &= \frac{\lim_{t\to\infty} \frac{A(t)}{t} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1,\, i\le I(t)} U_i(t)}{A(t)}}{\lim_{t\to\infty} \frac{I(t)}{t} \frac{\sum_{i=1}^{I(t)} U_i(t)}{I(t)}} \\
&= \frac{\lim_{t\to\infty} \frac{A(t)}{t} \overline{U_A}}{\lim_{t\to\infty} \frac{I(t)}{t} \overline{U}} \\
&\le \frac{\lambda p}{\lambda}, w.p.1 \qquad (45) \\
&= p, w.p.1 \qquad (46)
\end{aligned}$$

with equality in (45) if and only if the strategy Ψ_P satisfies Property 2 based on Proposition 1. Following (46), we have:

$$\lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1,\, i\le I(t)} U_i(t)}{t} = p \lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i(t)}{t}, w.p.1 \tag{47}$$

Then based on Equation (43) , (44), (45) and (47), we have:

$$\begin{aligned}
\lim_{t\to\infty} \frac{\mu T_{PP}(t)}{t} p &\overset{(44)}{=} \lim_{t\to\infty} \frac{\left(\sum_{i=1}^{I(t)} U_i(t) + \sum_{i=I(t)+1}^{\infty} U_i(t)\right) p}{t} \\
&= \lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i(t)}{t} p + \lim_{t\to\infty} \frac{\sum_{i=I(t)+1}^{\infty} U_i(t)}{t} p \qquad (48) \\
&\overset{(47),(45)}{\ge} \frac{\lim_{t\to\infty} \sum_{i\in\mathbb{Z}^+:R_i=1,\, i\le I(t)} U_i(t)}{t}, w.p.1 \qquad (49) \\
&\overset{(43)}{=} \lim_{t\to\infty} \frac{\mu T_{PS}(t)}{t}, w.p.1 \qquad (50)
\end{aligned}$$

with equality in (49) if and only if the strategy satisfies both Property 1 and Property 2. So if we put Equation (42) over t and take $t\to\infty$, we have:

$$\begin{aligned}
\lim_{t\to\infty} \frac{T_{PP}(t)}{t} &= \lim_{t\to\infty} \frac{T_{PS}(t)}{t} + \lim_{t\to\infty} \frac{T_{RI}(t)}{t} \\
&\le \lim_{t\to\infty} \frac{T_{PP}(t)}{t} p + \lim_{t\to\infty} \frac{T_{RI}(t)}{t}, w.p.1 \qquad (51) \\
(1-p) \lim_{t\to\infty} \frac{T_{PP}(t)}{t} &\le \lim_{t\to\infty} \frac{T_{RI}(t)}{t}, w.p.1 \qquad (52)
\end{aligned}$$

where (51) is from (50).

By replacing corresponding terms in Equation (52) with Equation (38) and (11), we have:

$$\alpha_{PP} \leq \frac{\alpha_{RI}}{1-p}, w.p.1 \tag{53}$$

and we know from fundamental queueing theory that:

$$\alpha_{RI} = 1 - \frac{\lambda ps}{\mu} \tag{54}$$

Then we have the result:

$$\alpha_{PP} \leq \frac{\mu - \lambda ps}{\mu(1-p)}, w.p.1 \tag{55}$$

And it follows that:

$$\alpha_{PR} = 1 - \alpha_{PP} \geq \frac{\lambda ps - \mu p}{\mu(1-p)}, w.p.1 \tag{56}$$

with equality in (55) and (56) if and only if the strategy satisfies both Property 1 and Property 2.

C PROOF OF COROLLARY 1

The average amount of proactive work done for each potential request in $\{i \in \mathbb{Z}^+ : i \leq I(t)\}$ by time t can be calculated by dividing the total amount of proactive work done for these requests by the total number $I(t)$, so

$$\begin{aligned}
\overline{U} &= \lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i}{I(t)} \\
&= \lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i(t)}{I(t)} \\
&= \lim_{t\to\infty} \frac{\sum_{i=1}^{\infty} U_i(t)}{I(t)} - \lim_{t\to\infty} \frac{\sum_{i=I(t)+1}^{\infty} U_i(t)}{I(t)} && (57) \\
&\leq \frac{\lim_{t\to\infty} \frac{\sum_{i=1}^{\infty} U_i(t)}{t}}{\lim_{t\to\infty} \frac{I(t)}{t}}, w.p.1 && (58) \\
&= \frac{\mu\alpha_{PP}}{\lambda}, w.p.1 && (59) \\
&\leq \frac{\mu - \lambda ps}{\lambda(1-p)}, w.p.1 && (60)
\end{aligned}$$

with equality in (58) and (60) if and only if the strategy satisfies both Property 1 and Property 2. We get (59) from (58) by the definition of α_{PP} and by the Strong Law of Large Numbers. The second term in (57) is 0 w.p.1 if and only if Ψ_P satisfies Property 1. Similarly, we have:

$$\begin{aligned}
\overline{S} &= \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\leq I(t)} S_i}{A(t)} \\
&= \frac{\lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\leq I(t)} S_i}{t}}{\lim_{t\to\infty} \frac{A(t)}{t}} && (61) \\
&= \frac{\mu\alpha_{PR}}{\lambda p}, w.p.1 && (62) \\
&\geq \frac{\lambda s - \mu}{\lambda(1-p)}, w.p.1 && (63)
\end{aligned}$$

with equality in (63) if and only if the strategy satisfies both Property 1 and Property 2. (62) is by the Strong Law of Large Numbers.

D PROOF OF PROPOSITION 2

Evolution of the Markov Chain: Consider the system starting from state $X_n = x_n$, $x_n \in \mathbb{Z}^+$ at τ_n. By Definition 3, it means that 1) $V\left(\tau_n^+\right) = 0$, 2) $U_{J(\tau_n^+)}\left(\tau_n^+\right) = 0$, and 3) $J\left(\tau_n^+\right) = P\left(\tau_n^+\right) + x_n$. From Condition 3), we know that the system starts proactively serving request $J\left(\tau_n^+\right) = P\left(\tau_n^+\right) + x_n$ right after τ_n. If the request $P\left(\tau_n^+\right) + x_n$ receives ϕ bits of proactive service before its arrival epoch $t_{P(\tau_n^+)+x_n}$, or an equivalent condition:

$$U_{P(\tau_n^+)+x_n}\left(t_{P(\tau_n^+)+x_n}\right) = \phi \tag{64}$$

is satisfied, it can be easily verified by Definition 3 that a transition happens right after the request $P\left(\tau_n^+\right) + x_n$ receives ϕ bits of proactive service. Therefore if (64) is satisfied, we have:

$$\tau_{n+1} < t_{P(\tau_n^+)+x_n} \tag{65}$$

By the definition of threshold-based strategies, an important fact is that:

$$\begin{aligned} \text{If } x_n > 1 : & U_i(\tau_n^+) = \phi, i = P(\tau_n^+) + 1, P(\tau_n^+) + 2, \ldots, P(\tau_n^+) + x_n - 1 \\ & U_i(\tau_n^+) = 0, i = P(\tau_n^+) + x_n, P(\tau_n^+) + x_n + 1, \ldots \\ \text{If } x_n = 1 : & U_i(\tau_n^+) = 0, i = P(\tau_n^+) + 1, P(\tau_n^+) + 2, \ldots \end{aligned} \tag{66}$$

given $X_n = x_n$. Therefore another equivalent condition to (64) is:

$$\begin{aligned} &\text{If } x_n = 1 : \phi \le \mu\left(t_{P(\tau_n^+)+x_n} - \tau_n\right) \\ &\text{If } x_n > 1 : \sum_{i=P(\tau_n^+)+1}^{P(\tau_n^+)+x_n-1} (s-\phi) R_i - V\left(t_{P(\tau_n^+)+x_n}\right) + \phi \le \mu\left(t_{P(\tau_n^+)+x_n} - \tau_n\right) \end{aligned} \tag{67}$$

$\sum_{i=P(\tau_n^+)+1}^{P(\tau_n^+)+x_n-1} (s-\phi) R_i$ represents the total amount of reactive work of all actual arrivals in $\left(\tau_n, t_{P(\tau_n^+)+x_n}\right)$. $V\left(t_{P(\tau_n^+)+x_n}\right)$ represents the amount of unfinished reactive work at the arrival epoch of request $P\left(\tau_n^+\right) + x_n$. Notice that $P\left(\tau_n^+\right) + x_n$ is the request being proactively worked on starting from τ_n. So the term $\sum_{i=P(\tau_n^+)+1}^{P(\tau_n^+)+x_n-1} (s-\phi) R_i - V\left(t_{P(\tau_n^+)+x_n}\right)$ represents the total amount of reactive work done in $\left(\tau_n, t_{P(\tau_n^+)+x_n}\right)$. The RHS means the total amount of work that can be done in $\left(\tau_n, t_{P(\tau_n^+)+x_n}\right)$. If Condition (67) is satisfied, it means that ϕ bits of proactive work can be done for request $J\left(\tau_n^+\right) = P\left(\tau_n^+\right) + x_n$ before $t_{P(\tau_n^+)+x_n}$, so (64) and (65) are satisfied. Next, we discuss the evolution of the Markov chain based on (67).

Case 1: If (67) is satisfied, (65) is true. In this case, we have $A_n < x_n$ and the following the transition happens:

$$\begin{aligned} X_{n+1} &= J\left(\tau_{n+1}^+\right) - P\left(\tau_{n+1}^+\right) \\ &= \left(J\left(\tau_n^+\right) + 1\right) - \left(P\left(\tau_n^+\right) + A_n\right) \\ &= x_n + 1 - A_n \end{aligned} \tag{68}$$

Case 2: If (67) is not satisfied, we know that no transition happens in $\left(\tau_n, t_{P(\tau_n^+)+x_n}\right)$. Because there have been x_n arrivals by $t^+_{P(\tau_n^+)+x_n}$, we have $A_n \ge x_n$. We are going to show that $X_{n+1} = 1$ in this case in the following.

Suppose we have $X_{n+1} \ge 2$, then we have $J\left(\tau_{n+1}^+\right) = P\left(\tau_{n+1}^+\right) + x_{n+1} > P\left(\tau_{n+1}^+\right) + 1$ by Definition 3. Then based on the definition of threshold-based strategies in Algorithm 1, there must an epoch $\tau' \in \left(t_{P(\tau_n^+)+X_n}, \tau_{n+1}\right)$ such that 1) $V\left(\tau'^+\right) = 0$, 2) $U_{J(\tau'^+)}\left(\tau'^+\right) = 0$, and 3)$J\left(\tau'^+\right) = P\left(\tau_{n+1}^+\right) + 1$.

Because we know $\tau' < \tau_{n+1}$, we have $J(\tau'^{+}) = P\left(\tau_{n+1}^{+}\right) + 1 \geq P(\tau'^{+}) + 1$. By Definition 3, a transition should happen at τ' which is earlier than τ_{n+1}. So a contradiction is achieved. Then if $A_n \geq x_n$:

$$X_{n+1} = 1 \tag{69}$$

By summarizing Cases 1 and 2, we have

$$X_{n+1} = \max\{X_n + 1 - A_n, 1\}, \forall n = 0, 1, \ldots \tag{70}$$

Proof of Markovian Property: Now we consider (67). Condition (67) is determined by X_n, $\left\{R_i : i > P\left(\tau_n^{+}\right), i \in \mathbb{Z}^{+}\right\}$, $\left\{t_i : i > P\left(\tau_n^{+}\right), i \in \mathbb{Z}^{+}\right\}$ and $V\left(t_{P(\tau_n^{+})+x_n}\right)$. If the realization (i.e., R_i's), arrival epoch (i.e., t_i's) and the amount of reactive work to be done of each actual arrival are determined, the term $V\left(t_{P(\tau_n^{+})+x_n}\right)$ is also deterministic. $\left\{R_i : i > P\left(\tau_n^{+}\right), i \in \mathbb{Z}^{+}\right\}$ are IID Binomial random variables which are memoryless. $\left\{t_i : i > P\left(\tau_n^{+}\right), i \in \mathbb{Z}^{+}\right\}$ are determined by the Poisson process $\{P(t); t > 0\}$, which are also memoryless. The amount of reactive work to be done of each actual arrival is determined by X_n and the arrival processes after τ_n. Therefore X_{n+1} only depends on X_n and what happens after τ_n, and the chain is Markovian by definition.

E PROOF OF PROPOSITION 3

Recall the definition of $p_k^{\phi} \triangleq Pr\{A_n = k | X_n = x_n\}, x_n > k, k = 0, 1, 2, \ldots$. In order to calculate the transition probabilities $\left\{p_k^{\phi}, k = 0, 1, 2, \ldots\right\}$, we consider the probabilities $p_k^{\phi} = Pr\{A_n = k | X_n = \infty\}$, $k = 0, 1, 2, \ldots$, based on Fact (19). p_k^{ϕ} can then be interpreted as the probability that there are k potential arrivals before the next transition happens given $X_n = \infty$. Then the target term $\sum_{k=0}^{\infty} p_k^{\phi}(1-k)$ can be explained as the expected drift of the next transition, i.e., $E[X_{n+1} - X_n | X_n = \infty]$, in the Markov chain. In the following, we are going to compute the probabilities of $\left\{p_k^{\phi}, k = 0, 1, 2 \ldots\right\}$, with respect to different values of ϕ.

Distributions of T_n and A_n: We first analyze the distribution of $T_n | X_n = \infty$ and $A_n | X_n = \infty$. Inspired by the methods used in the analysis of the distribution of busy periods in M/G/1 queues in Section 8-4 of [20], we use a similar method.

Define function $T(\omega_1, \omega, \lambda, p)$ as the length of a time interval starting from the arrival epoch of the first job in an empty system, to the epoch when the system becomes empty for the first time again. The arrivals follow a Poisson process with an overall arrival rate of λ, where each arrival is realized with probability p, IID. The service time of the first job is ω_1, and the service time of the next arrivals is ω if realized.

Notice that queueing disciplines will not affect the length of this time interval, as long as the system is work-conserving. Specifically, if we select $\omega_1 = \frac{\phi}{\mu}$ and $\omega = \frac{s-\phi}{\mu}$, we have $T\left(\frac{\phi}{\mu}, \frac{s-\phi}{\mu}, \lambda, p\right) = (T_n | X_n = \infty)$. If we select $\omega_1 = \frac{s-\phi}{\mu}$ and $\omega = \frac{s-\phi}{\mu}$, $T\left(\frac{s-\phi}{\mu}, \frac{s-\phi}{\mu}, \lambda, p\right)$ is the length of the time interval from the arrival epoch of an actual request to the epoch it gets completely served under Last-In-First-Out (LIFO) discipline. It is also the length of a busy period in our proposed system given $X_n = \infty$, which is the time interval from the arrival of the first actual request when $V(t) = 0$, to the epoch when $V(t) = 0$ again.

Denote the number of potential arrivals when $V(t) = 0$ during $T(\omega_1, \omega, \lambda, p)$ as $N_P \sim \mathcal{P}(\lambda\omega_1)$, where $\mathcal{P}(\cdot)$ is the Poisson distribution. Notice that N_p is different from the number of arrivals in $T(\omega_1, \omega, \lambda, p)$ because some arrivals happen when the server is working reactively, i.e.$V(t) > 0$. Denote the number of actual arrivals among these N_p arrivals as $N_A \sim \mathcal{B}(N_P, p)$, where $\mathcal{B}(\cdot, \cdot)$ is

the Binomial distribution. When an actual request among N_A arrives, a busy period starts. The length of each busy period follows the distribution $T(\omega, \omega, \lambda, p)$, IID.

First we derive $T(\omega, \omega, \lambda, p)$. Following similar arguments in Section 8-4 of [20], we consider LIFO queueing discipline for unfinished work $V(t)$, which does not affect the length of the time interval $T(\omega, \omega, \lambda, p)$. Then we have:

$$E[T(\omega, \omega, \lambda, p) | N_P, N_A] = \omega + N_A E[T(\omega, \omega, \lambda, p)] + (N_P - N_A)\,0 \tag{71}$$

Then by definitions of N_A and N_P, we have:

$$E[T(\omega, \omega, \lambda, p) | N_P] = \omega + pN_P E[T(\omega, \omega, \lambda, p)] \tag{72}$$

$$E[T(\omega, \omega, \lambda, p)] = \omega + p\lambda\omega E[T(\omega, \omega, \lambda, p)] \tag{73}$$

So we have:

$$E[T(\omega, \omega, \lambda, p)] = \frac{\omega}{1 - p\lambda\omega} \tag{74}$$

Similarly, we can derive $E[T(\omega_1, \omega, \lambda, p)]$. We know the service time of the first job is ω_1, and each busy period follows $T(\omega, \omega, \lambda, p)$, we have:

$$E[T(\omega_1, \omega, \lambda, p)] = \omega_1 + \lambda p \omega_1 E[T(\omega, \omega, \lambda, p)] \tag{75}$$

$$= \omega_1 + \lambda p \omega_1 \frac{\omega}{1 - p\lambda\omega} \tag{76}$$

By replacing corresponding terms, we have

$$E[T_n | X_n = \infty] = E\left[T\left(\frac{\phi}{\mu}, \frac{s-\phi}{\mu}, \lambda, p\right)\right]$$

$$= \frac{\phi}{\mu} + \lambda p \frac{\phi}{\mu} \frac{\frac{s-\phi}{\mu}}{1 - p\lambda\frac{s-\phi}{\mu}} \tag{77}$$

$$= \frac{\phi}{\mu - p\lambda(s-\phi)} \tag{78}$$

$$= \frac{1}{\frac{\mu - \lambda p s}{\phi} + p\lambda} \tag{79}$$

Similarly, we can define $A(\omega_1, \omega, \lambda, p)$ as the number of arrivals in the next transition given $X_n = \infty$. With similar arguments, we have:

$$E[A(\omega, \omega, \lambda, p)] = \frac{\lambda\omega}{1 - \lambda\omega p} \tag{80}$$

$$E[A(\omega_1, \omega, \lambda, p)] = \lambda p \omega_1 \frac{\lambda\omega}{1 - \lambda\omega p} + \lambda\omega_1 \tag{81}$$

$$E[A_n | X_n = \infty] = E\left[A\left(\frac{\phi}{\mu}, \frac{s-\phi}{\mu}, \lambda, p\right)\right] = \frac{\lambda}{\frac{\mu - \lambda p s}{\phi} + p\lambda} \tag{82}$$

An interesting fact is that if we choose ϕ according to U^*, as defined in (15), we have

$$E[A_n | X_n = \infty] \begin{cases} < 1, if & \phi < U^* \\ = 1, if & \phi = U^* \\ > 1, if & \phi > U^* \end{cases} \tag{83}$$

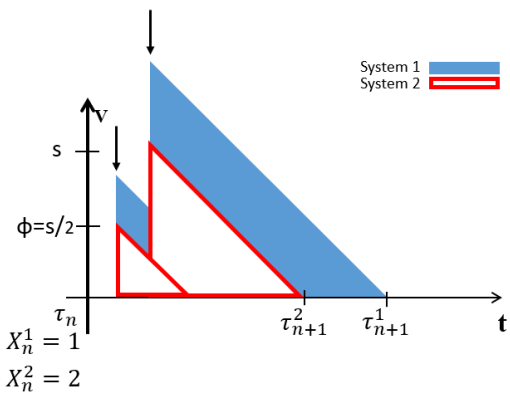

Fig. 14. Comparison of System 1 and System 2 in the Proof of Proposition 4

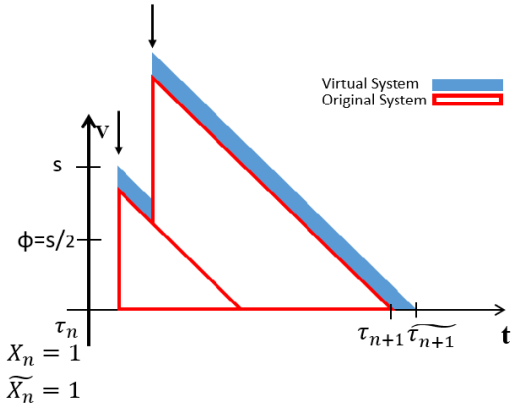

Fig. 15. Comparison of the Proactive System and the Virtual System in the Proof of Proposition 4

Notice that $Pr\{A_n = k|X_n = \infty\} = p_k^\phi, \forall k = 0, 1, \ldots$, and

$$E[A_n|X_n = \infty] = \sum_{k=0}^{\infty} p_k^\phi k \tag{84}$$

So we have

$$\sum_{k=0}^{\infty} p_k^\phi k \begin{cases} < 1, if & \phi < U^* \\ = 1, if & \phi = U^* \\ > 1, if & \phi > U^* \end{cases} \tag{85}$$

And our conclusion directly follows:

$$\sum_{k=0}^{\infty} p_k^\phi (1-k) \begin{cases} > 0, if & \phi < U^* \\ = 0, if & \phi = U^* \\ < 0, if & \phi > U^* \end{cases} \tag{86}$$

F PROOF OF PROPOSITION 4

In order to prove $E[T_n|X_n = x_n] < \infty,\ \forall x_n \in \mathbb{Z}^+$, we first prove that:

$$E[T_n|X_n = 1] \geq E[T_n|X_n = k], \forall k > 1 \tag{87}$$

then prove that:

$$E[T_n|X_n = 1] < \infty \tag{88}$$

to finish the proof.

Proof of (87):

First, we prove the following:

$$E[T_n|X_n = 1] \geq E[T_n|X_n = k], \forall k > 1 \tag{89}$$

Consider two systems under Ψ_P^ϕ which start from τ_n with $P_1(\tau_n) = P_2(\tau_n)$, but in different states : $X_n^1 = 1$ in the first system and $X_n^2 = k, k > 1$ in the second system. Based on (66), no proactive service has been done for any future requests at τ_n in the first system, and the first $k-1$ future

requests have received ϕ bits of proactive service by time τ_n in the second system. Recall that $J(t)$ denotes the request the server would proactively work on if the $V(t) = 0$ at t. Here we use $J_1(t)$ for the first system and $J_2(t)$ for the second system.

Because we assume $P_1(\tau_n) = P_2(\tau_n)$, $X_n^1 = 1$ and $X_n^2 = k, k > 1$, we have $J_2\left(\tau_n^+\right) > J_1\left(\tau_n^+\right)$ by Definition 3. Then if we consider the same arrival processes after τ_n in both systems under the same strategy Ψ_P^ϕ, we have

$$J_2(t) \geq J_1(t), \forall t \geq \tau_n \tag{90}$$

Then we have $\tau_{n+1}^1 \geq \tau_{n+1}^2$ by Definition 3, which means a transition always happens in the second system no later than the first system. Therefore we have:

$$T_n | X_n = 1 \geq T_n | X_n = k, k \geq 2 \tag{91}$$

It is true for every sample path, so we have:

$$E[T_n | X_n = 1] \geq E[T_n | X_n = k], \forall k > 1 \tag{92}$$

An example of the comparison can be found in Figure 14.

Proof of (88): Next, we prove that $E[T_n | X_n = 1] < \infty$. Again, we use the method of comparisons to prove it.

We compare the proactive system with a virtual system. Both systems start from state $X_n = 1$ ($\tilde{X_n} = 1$ in the virtual system) at τ_n. In the virtual system, the server stops proactively serving any requests from τ_n. Our goal is to find the earliest epoch $\tau^* > \tau_n$ which satisfies:

$$1)\tilde{V}\left(\tau^{*+}\right) = 0, 2)\tilde{U}_{\tilde{j}(\tau_n^+)}(\tau_n^+) = 0, 3)\tilde{P}\left(\tau^{*+}\right) = \tilde{I}\left(\tau^{*+}\right) > I\left(\tau_n^+\right) \tag{93}$$

Note that these conditions are very similar to the conditions in Definition 3. We consider τ^* as the next transition time in the virtual system. Correspondingly, we define $\tilde{T_n} = \tau^* - \tau_n$. Note that $\tilde{P}(\tau^{*+}) = \tilde{I}(\tau^{*+}) > I\left(\tau_n^+\right)$ is a stronger condition to Condition 3) in Definition 3. Based on the definitions, we are going to prove

$$\left(\tilde{T_n} | \tilde{X_n} = 1\right) \geq (T_n | X_n = 1) \tag{94}$$

in two systems under the same sample path.

Now we consider the same sample path in both the proactive system under Ψ_P^ϕ and the virtual system starting from $X_n = 1$ and $\tilde{X_n} = 1$ from τ_n. An example of the comparison is shown in Figure 15. Since no proactive work will be done in the virtual system before τ^*, all the actual arrivals need to receive s bits reactively in the virtual system, which is no fewer than the proactive system for each request. Therefore similar to the previous arguments we did for Equation (90), we have:

$$\tilde{J}(t) \leq J(t), \forall t > \tau_n \tag{95}$$

If we compare the conditions in (93) with the conditions in Defition 3, we can see that (93) is a stronger condition because if $P(\tau^{*+}) = I(\tau^{*+})$, we must have $\tilde{J}(\tau^{*+}) \geq I(\tau^{*+}) + 1 > P(\tau^{*+})$. Therefore the transition in the proactive system will happen no later than the virtual system. Therefore the transition time $\left(\tilde{T_n} | \tilde{X_n} = 1\right) \geq (T_n | X_n = 1)$ is true along every sample path.

Construction of τ^*: Here we aim to find the epoch τ^* in the virtual system which satisfies conditions in Equation (93).

Our target is to find a busy period which starts with one actual arrival, and no other potential arrivals happen before it ends. The epoch τ^* can be found when such a busy period ends because: 1) the server becomes idle so $\tilde{V}(\tau^{*+}) = 0$, 2) $\tilde{U}_{\tilde{j}(\tau_n^+)}(\tau_n^+) = 0$ is always true in the virtual system

based on its assumption, 3) the latest arrival is an actual arrival so $\tilde{P}(\tau^{*+}) = \tilde{I}(\tau^{*+})$. Because there should be at least one actual arrival after τ_n, we have $\tilde{I}(\tau^{*+}) > I\left(\tau_n^+\right)$ so Condition 3) is satisfied.

Recall that we assume $\lambda sp < \mu$, so the virtual system is stable. The expected idle period length in the virtual system is then

$$E[I_V] = \frac{1}{\lambda p} \tag{96}$$

where I_V is defined as the length of an idle period in the virtual system. Based on $\frac{E[B_V]}{E[B_V]+E[I_V]} = \rho = \frac{\lambda ps}{\mu}$, where B_V is the length of a busy period in the virtual system, we can also calculate the expected length of a busy period in the virtual system $E[B_V]$. So we know that $E[I_V] < \infty, E[B_V] < \infty$.

The next step is to find such a busy period. Every time a busy period starts with an actual arrival, the probability that there are no potential arrivals during the service time of the actual arrival $\frac{s}{\mu}$ is $e^{-\lambda\frac{s}{\mu}}$, IID. Therefore, the expected number of busy periods that such a busy period happens for the first time is $E[N_B] = \frac{1}{e^{-\lambda\frac{s}{\mu}}} = e^{\lambda\frac{s}{\mu}}$, where N_B is the number of busy periods when the first busy period satisfying the condition is observed. So the expected time that such a busy period happens is then bounded as:

$$E\left[\tilde{T_n}|\tilde{X_n} = 1\right] \leq E[N_B](E[I_V] + E[B_V]) < \infty \tag{97}$$

Define the bound $E_V \triangleq E[N_B](E[I_V] + E[B_V])$, which is a deterministic finite number given system parameters λ, p, s, μ. Therefore we have our bound on $E[T_n|X_n]$:

$$E[T_n|X_n] \leq E[T_n|X_n = 1] \leq E\left[\tilde{T_n}|\tilde{X_n} = 1\right] = E_V < \infty, \ \forall X_n \in \mathbb{Z}^+ \tag{98}$$

So we proved $E[T_n|X_n = k] < \infty, \forall k \in \mathbb{Z}^+$ by combining (87) and (88).
Similarly we can prove $E[A_n|X_n = k] < \infty, \forall k \in \mathbb{Z}^+$.

G PROOF OF LEMMA 1

First assume that by time t, there have been $M(t) \triangleq \max\{m|\tau_m \leq t\}$ transitions in the Markov chain under Ψ_P^ϕ. Then we have the following inequalities of $M(t)$:

$$M(t) \leq P(t) + \frac{\mu t}{\phi} \tag{99}$$

$$\lim_{t\to\infty} \frac{t}{M(t)} \leq E_V, w.p.1 \tag{100}$$

where $E_V \triangleq E[N_B](E[I_V] + E[B_V])$ is the bound in (97). Equation (99) is by the fact that a transition happens either when the server finishes proactively serving a request with ϕ bits, or some potential arrival happens before it receives ϕ bits of proactive service. The term $\lim_{t\to\infty} \frac{t}{M(t)}$ is the limiting average of T_n. Equation (100) is from Proposition 4. Take (99) over t and take limit of $t \to \infty$, we get:

$$\lim_{t\to\infty} \frac{M(t)}{t} \leq \lim_{t\to\infty} \left(\frac{P(t)}{t} + \frac{\mu}{\phi}\right) = \lambda + \frac{\mu}{\phi}, w.p.1 \tag{101}$$

Combining it with (100), we have:

$$\lambda + \frac{\mu}{\phi} \geq \lim_{t\to\infty} \frac{M(t)}{t} \geq \frac{1}{E_V} \tag{102}$$

On the other hand, recall that if $J(t) > P(t)$, we have (66). Based on Definition 3, $X_{n+1} - X_n \leq 1, \forall n = 0, 1, \ldots$. Then we have:

$$J(t) - P(t) \leq \max\left\{X_{M(t)}, X_{M(t)+1}\right\} \leq X_{M(t)} + 1 \tag{103}$$

$$J(t) - P(t) \geq \min\left\{X_{M(t)}, X_{M(t)+1}\right\} \geq X_{M(t)+1} - 1 \tag{104}$$

$$\forall t \in \left(\tau_{M(t)}, \tau_{M(t)+1}\right)$$

Therefore we have $\forall t \in (\tau_n, \tau_{n+1})$:

$$\sum_{i=P(t)+1}^{\infty} U_i(t) \leq \max\left\{\phi\left(J(t) - P(t)\right), 0\right\} \tag{105}$$

$$\leq \phi\left(X_{M(t)} + 1\right) \tag{106}$$

$$\sum_{i=P(t)+1}^{\infty} U_i(t) \geq \max\left\{\phi\left(J(t) - P(t) - 1\right), 0\right\} \tag{107}$$

$$\geq \phi\left(X_{M(t)+1} - 2\right) \tag{108}$$

$$, \forall t \in \left(\tau_{M(t)}, \tau_{M(t)+1}\right)$$

(105) is achieved by considering the amount of proactive service done for request $J(t)$ as ϕ. (106) is from (103). (107) is achieved by considering the amount of proactive service done for request $J(t)$ as 0, and (108) is from (104). Therefore for all t, we have:

$$\frac{\sum_{i=P(t)+1}^{\infty} U_i(t)}{t} \leq \frac{\phi\left(X_{M(t)} + 1\right)}{t} \tag{109}$$

$$\frac{\sum_{i=P(t)+1}^{\infty} U_i(t)}{t} \geq \frac{\phi\left(X_{M(t)+1} - 2\right)}{t} \tag{110}$$

If we take the limit of $t \to \infty$ we have

$$\begin{aligned}\lim_{t\to\infty} \frac{\sum_{i=P(t)+1}^{\infty} U_i(t)}{t} &\leq \lim_{t\to\infty} \frac{\phi\left(X_{M(t)} + 1\right)}{t} \\ &= \lim_{t\to\infty} \frac{\phi\left(X_{M(t)} + 1\right)}{M(t)} \frac{M(t)}{t} \\ &\leq \lim_{n\to\infty} \frac{X_n}{n} \phi\left(\lambda + \frac{\mu}{\phi}\right)\end{aligned} \tag{111}$$

$$\begin{aligned}\lim_{t\to\infty} \frac{\sum_{i=P(t)+1}^{\infty} U_i(t)}{t} &\geq \lim_{t\to\infty} \frac{\phi\left(X_{M(t)+1} - 2\right)}{t} \\ &= \lim_{t\to\infty} \frac{\phi\left(X_{M(t)+1} - 2\right)}{M(t)} \frac{M(t)}{t} \\ &\geq \lim_{n\to\infty} \frac{X_n}{n} \frac{\phi}{E_V}\end{aligned} \tag{112}$$

So if we know $\lim_{n\to\infty} \frac{X_n}{n} = 0, w.p.1$, we have $\lim_{t\to\infty} \frac{\sum_{i=P(t)+1}^{\infty} U_i(t)}{t} = 0, w.p.1$ from (111). And if $\lim_{t\to\infty} \frac{\sum_{i=P(t)+1}^{\infty} U_i(t)}{t} = 0, w.p.1$, we have $\lim_{n\to\infty} \frac{X_n}{n} = 0, w.p.1$ from (112). So by Definition 1, the threshold-based strategy Ψ_P^{ϕ} satisfies Property 1 if and only if the corresponding Markov chain satisfies $\lim_{n\to\infty} \frac{X_n}{n} = 0, w.p.1$.

H PROOF OF THEOREM 3

Case 1: If $\phi < U^*$, we know that the chain is transient from Theorem 2. Therefore, $\exists N > 0$ such that:

$$X_n > 1, \forall n > N$$

with probability 1. From the Nth transition, we look at the drifts, i.e. $\Delta_m \triangleq X_{m+1} - X_m, \forall m \in \mathbb{Z}^+$. Then for all $n > N$:

$$\begin{aligned} X_n &= X_N + \sum_{i=N}^{n-1} (X_{i+1} - X_i) \\ &= X_N + \sum_{i=N}^{n-1} \Delta_i \\ &= X_N + \sum_{k=0}^{\infty} \sum_{i \in \mathbb{Z}^+ : \Delta_i = 1-k, N \le i < n} \Delta_i \qquad (113) \\ &= X_N + \sum_{k=0}^{\infty} (1-k) \frac{|\{i \in \mathbb{Z}^+ : \Delta_i = 1-k, N \le i < n\}|}{n-N} (n-N) \qquad (114) \end{aligned}$$

(113) is achieved by grouping transitions based on the size of drifts. Notice that $\frac{|\{i \in \mathbb{Z}^+ : \Delta_i = 1-k, N \le i < n\}|}{n-N}$ is the fraction of drifts with value $1-k$, where there are k arrivals before the next transition. Because the chain never revisits state 1 after N, the probability that there are k arrivals is p_k^ϕ. So as $n \to \infty$, we have:

$$\lim_{n\to\infty} \frac{|\{i \in \mathbb{Z}^+ : \Delta_i = 1-k, N \le i < n\}|}{n-N} = p_k^\phi, w.p.1 \qquad (115)$$

based on Strong Law of Large Numbers. If we take both sides over n and take the limit of $n \to \infty$, we have:

$$\begin{aligned} \lim_{n\to\infty} \frac{X_n}{n} &= \lim_{n\to\infty} \left(\frac{X_N}{n} + \sum_{k=0}^{\infty} (1-k) \frac{|\{i \in \mathbb{Z}^+ : \Delta_i = 1-k, N \le i < n\}|}{n-N} \frac{(n-N)}{n} \right) \\ &= \sum_{k=0}^{\infty} (1-k) \lim_{n\to\infty} \left(\frac{|\{i \in \mathbb{Z}^+ : \Delta_i = 1-k, N \le i < n\}|}{n-N} \frac{(n-N)}{n} \right) \\ &= \sum_{k=0}^{\infty} (1-k) p_k^\phi \\ &> 0 \qquad (116) \end{aligned}$$

based on Proposition 3. So when $\phi < U^*$, the threshold-based strategy does not satisfy Property 1 based on Lemma 1.

Case 2: If $\phi \ge U^*$, we consider a virtual strategy. In this strategy, the server can do proactive work at the rate of

$$\mu_\epsilon \triangleq \frac{\frac{\lambda}{\frac{\mu-\lambda ps}{\phi}+\lambda p}}{1-\epsilon} \mu \ge \frac{\frac{\lambda}{\frac{\mu-\lambda ps}{U^*}+\lambda p}}{1-\epsilon} \mu = \frac{1}{1-\epsilon} \mu > \mu, \epsilon \in (0,1) \qquad (117)$$

In this case, define $J_\epsilon(t)$ as the request that the system would proactively work on at time t under the virtual strategy. Because the system is always working at a strictly higher rate $\mu_\epsilon > \mu$ with the

virtual strategy, we have $J_\epsilon(t) \geq J(t), \forall t$ if the two systems are under the same sample path. Then by Definition 3, we have:

$$X_n^\epsilon \geq X_n, \forall n \in \mathbb{Z}^+ \tag{118}$$

where X_n^ϵ is defined as the states under the virtual strategy.

Following the same steps of Proposition 3, we can derive the new set of transition probabilities $\left\{p_k^{\epsilon\phi}\right\}$, and prove that the Markov chain under the virtual strategy is transient. Specifically:

$$\lim_{n\to\infty} \frac{X_n^\epsilon}{n} = \sum_{k=0}^{\infty} p_k^{\epsilon\phi}(1-k) = \epsilon, \forall \epsilon \in (0,1) \tag{119}$$

So $\forall \epsilon \in (0,1)$, we have

$$\lim_{n\to\infty} \frac{X_n}{n} \leq \lim_{n\to\infty} \frac{X_n^\epsilon}{n} = \epsilon \tag{120}$$

And if we take $\epsilon \to 0$:

$$\lim_{n\to\infty} \frac{X_n}{n} \leq \lim_{\epsilon\to 0} \lim_{n\to\infty} \frac{X_n^\epsilon}{n} = 0 \tag{121}$$

Therefore based on Lemma 1, the threshold-based strategy satisfies Property 1 when $\phi \geq U^*$.

Then by summarizing Cases 1 and 2, the threshold-based strategy Ψ_P^ϕ satisfies Property 1 if and only if $\phi \geq U^*$.

I PROOF OF THEOREM 4

In order to prove Theorem 4, we first consider the following Lemma 2. The idea of Lemma 2 is to look at the proactive work done within one transition. Under strategy Ψ_P^ϕ, denote the total amount of proactive work done in (τ_n, τ_{n+1}) for all potential requests as ζ_n, and denote the amount of proactive work done in (τ_n, τ_{n+1}) for actual requests as ζ_n^A. We investigate the expectation of ζ_n and ζ_n^A conditioned on X_n and X_{n+1} in Lemma 2.

Lemma 2.

$$E\left[\zeta_n^A | X_n = k, X_{n+1} = l\right] = E\left[\zeta_n | X_n = k, X_{n+1} = l\right] p, \ if\ l > 1 \tag{122}$$

$$E\left[\zeta_n^A | X_n = k, X_{n+1} = l\right] \leq E\left[\zeta_n | X_n = k, X_{n+1} = l\right] p, \ if\ l = 1 \tag{123}$$

$$E\left[\zeta_n^A | X_n = k, X_{n+1} = l\right] < E\left[\zeta_n | X_n = k, X_{n+1} = l\right] p, \ if\ k = 1, l = 1 \tag{124}$$

Proof. **Proof of (122):** Given the starting state $X_n = k$ at τ_n, we focus on the request $P(\tau_n) + k$ which is the request that starts to receive proactive service from τ_n.

If $X_{n+1} > 1$, it means that the server is able to proactively serve request $P(\tau_n) + k$ before it arrives, i.e. $t_{P(\tau_n)+k} > \tau_{n+1}$. So we have the following:

$$(\zeta_n | X_n = k, X_{n+1} = l) = \phi \tag{125}$$

$$\left(\zeta_n^A | X_n = k, X_{n+1} = l\right) = \begin{cases} \phi, & if\ R_{P(\tau_n)+i} = 1 \\ 0, & if\ R_{P(\tau_n)+i} = 0 \end{cases}, \forall k \geq 1, \forall l > 1 \tag{126}$$

Because $Pr\left\{R_{P(\tau_n)+k} = 1\right\} = p$ independently, we have:

$$E\left[\zeta_n^A | X_n = k, X_{n+1} = l\right] = E\left[\zeta_n | X_n = k, X_{n+1} = l\right] p = \phi p, \ \forall k \geq 1, \forall l > 1 \tag{127}$$

So (122) is proved.

Proof of (123): If $X_{n+1} = 1$, which means that the request $P(\tau_n) + k$ arrives before it receives ϕ bits of proactive service, i.e. $t_{P(\tau_n)+k} < \tau_{n+1}$, we know that

(1) All the proactive work done in $(\tau_n, t_{P(\tau_n)+k})$ is for request $P(\tau_n) + k$;
(2) All the proactive work done in $(t_{P(\tau_n)+k}, \tau_{n+1})$ are for requests that are not realized.

Both of these facts are by the definition of threshold-based strategy. Because the server will keep proactively serving request $P(\tau_n)+k$ until it receives ϕ bits proactively or until it arrives, statement (1) is true. For statement (2), if any request that has not arrived starts to be proactively served, a transition should happen at the moment it starts receiving proactive service by Definition 3. Therefore before the transition happens, i.e. τ_{n+1}, there should be no proactive service for future potential arrivals. Take what happens in (τ_3, τ_4) in Figure 4 as an example. The server starts proactively serving request 4 at τ_3. In (τ_3, t_4), all the proactive work are done for request 4. All the proactive work in (t_4, τ_4) are done for the requests which are not realized.

Based on the discussions above, we have the following analysis. Consider the system starting at τ_n from state $X_n = k \in \mathbb{Z}^+$. Define a tuple of random vectors $\Theta_{n,k} \triangleq (\xi_{n,k}, \nu_{n,k})$, where $\xi_{n,k} \triangleq (t_{P(\tau_n)+1}, t_{P(\tau_n)+2}, \ldots, t_{P(\tau_n)+k})$ denotes a random vector of the next k arrival epochs t_i's after τ_n, and $\nu_{n,k} \triangleq (R_{P(\tau_n)+1}, R_{P(\tau_n)+2}, \ldots, R_{P(\tau_n)+k-1})$ denotes a random vector of the next $k-1$ R_i's after τ_n. A realization $\Theta_{n,k} = \theta_{n,k}$ determines a set of sample paths after τ_n, where the first k arrival epochs and the realization of the first $k-1$ arrivals are determined. Given $X_n = k$ and $\Theta_{n,k} = \theta_{n,k}$, what happens in the system during $(\tau_n, t_{P(\tau_n)+k})$ is deterministic. We also know whether $\tau_{n+1} > t_{P(\tau_n)+k}$ or not, which determines if $X_{n+1} = 1$ or $X_{n+1} > 1$, as discussed in the proof of Proposition 2.

Define $Q_{n,k,1}$ as:

$$Q_{n,k,1} \triangleq \{\theta_{n,k} : X_{n+1} = 1, X_n = k\} \tag{128}$$

which represents the set of sample paths under which the system transits from state k to 1 starting from τ_n.

Then $\forall \theta_{n,k} \in Q_{n,k,1}, \forall k \in \mathbb{Z}^+, n = 0, 1, \ldots$ we have:

$$(\zeta_n | \Theta_{n,k} = \theta_{n,k}, X_n = k) \geq (U_{P(\tau_n)+k}(t_{P(\tau_n)+k}) | \Theta_{n,k} = \theta_{n,k}, X_n = k) \tag{129}$$

$$\left(\zeta_n^A | \Theta_{n,k} = \theta_{n,k}, X_n = k\right) = \begin{cases} (U_{P(\tau_n)+k}(t_{P(\tau_n)+k}) | \Theta_{n,k} = \theta_{n,k}, X_n = k), & if\ R_{P(\tau_n)+k} = 1 \\ 0, & if\ R_{P(\tau_n)+k} = 0 \end{cases} \tag{130}$$

Therefore $\forall \theta_{n,k} \in Q_{n,k,1}, \forall k \in \mathbb{Z}^+, n = 0, 1, \ldots$:

$$\begin{aligned} E_R[\zeta_n^A | \Theta_{n,k} = \theta_{n,k}, X_n = k] &= (U_{P(\tau_n)+k}(t_{P(\tau_n)+k}) | \Theta_{n,k} = \theta_{n,k}, X_n = k)\, p \\ &\leq E_R[\zeta_n | \Theta_{n,k} = \theta_{n,k}, X_n = k] p \end{aligned} \tag{131}$$

where $E_R[\cdot]$ means expectation with respect to $R_{P(\tau_n)+k}$. And by definition of $Q_{n,k,1}$, we have:

$$\begin{aligned} E[\zeta_n^A | X_n &= k, X_{n+1} = 1] \\ &= E[\zeta_n^A | X_n = k, \Theta_{n,k} \in Q_{n,k,1}] \\ &= \int_{Q_{n,k,1}} Pr\{\Theta_{n,k} = \theta_{n,k} | \Theta_{n,k} \in Q_{n,k,1}\} \cdot E_R[\zeta_n^A | \Theta_{n,k} = \theta_{n,k}, X_n = k] d\theta_{n,k} \\ &\leq \int_{Q_{n,k,1}} Pr\{\Theta_{n,k} = \theta_{n,k} | \Theta_{n,k} \in Q_{n,k,1}\} \cdot E_R[\zeta_n | \Theta_{n,k} = \theta_{n,k}, X_n = k] p d\theta_{n,k} \\ &= E[\zeta_n | X_n = k, X_{n+1} = 1]\, p \end{aligned} \tag{132}$$

where (132) is by replacing corresponding terms according to (131). So we have:

$$E\left[\zeta_n^A | X_n = k, X_{n+1} = 1\right] \leq E\left[\zeta_n | X_n = k, X_{n+1} = 1\right] p, \forall k \in \mathbb{Z}^+ \tag{133}$$

so (123) is proved.

Proof of (124): The system starts from state $X_n = 1$. It means at time τ_n, no proactive work is done for any of the potential requests which have not arrived yet. Similar to the method we used to prove (123), we focus on the set $Q_{n,1,1}$ in this case. Recall that $\theta_{n,1} \in Q_{n,1,1}$ if and only if $X_{n+1} = 1$ given $X_n = 1$ and $\Theta_{n,1} = \theta_{n,1}$. Notice that $\theta_{n,1} = (\xi(n,1), \nu(n,1))$ where $\xi(n,1) = \left(t_{P(\tau_n)+1}\right)$ and $\nu(n,1)$ is an empty vector, which means the arrival epoch of request $P(\tau_n)+1$ determines whether $X_{n+1} = 1$ or not. To be specific, $X_{n+1} = 1$ if and only if $t_{P(\tau_n)+1} < \frac{\phi}{\mu} + \tau_n$. So:

$$\begin{aligned} Q_{n,1,1} &= \left\{\theta_{n,1} : X_{n+1} = 1, X_n = 1\right\} \\ &= \left\{\theta_{n,1} : t_{P(\tau_n)+1} < \frac{\phi}{\mu} + \tau_n\right\} \end{aligned} \tag{134}$$

We consider another set of sample paths $\overline{Q_{n,1,1}}$ which is defined as:

$$\overline{Q_{n,1,1}} \triangleq \left\{\theta_{n,1} : t_{P(\tau_n)+1} < \tau_n + \frac{\phi}{2\mu}, t_{P(\tau_n)+2} - t_{P(\tau_n)+1} \geq \frac{\phi}{2\mu}\right\} \tag{135}$$

By comparing (135) and (134), it is true that $\overline{Q_{n,1,1}} \subsetneq Q_{n,1,1}$.

We discuss the value of ζ_n under condition $\Theta_{n,1} = \theta_{n,1} \in \overline{Q_{n,1,1}}$. If $R_{P(\tau_n)+1} = 1$, the system will proactively work on request $P(\tau_n)+1$ until $t_{P(\tau_n)+1}$. Consequently, request $P(\tau_n)+1$ receives fewer than $\frac{\phi}{2}$ bits of proactive service due to the definition of $\overline{Q_{n,1,1}}$. If $R_{P(\tau_n)+1} = 0$, the system will proactively work on request $P(\tau_n)+1$, until it receives ϕ bits from proactive service or until $t_{P(\tau_n)+2}$. Therefore we have $\forall \theta_{n,1} \in \overline{Q_{n,1,1}}, \forall n = 0, 1, \ldots$:

$$\begin{aligned} &\left(\zeta_n | \Theta_{n,1} = \theta_{n,1}, X_n = 1\right) \\ &\quad \geq \begin{cases} \left(U_{P(\tau_n)+1}\left(t_{P(\tau_n)+1}\right) | \Theta_{n,1} = \theta_{n,1}, X_n = 1\right), & if\ R_{P(\tau_n)+1} = 1 \\ \left(U_{P(\tau_n)+1}\left(t_{P(\tau_n)+1}\right) | \Theta_{n,1} = \theta_{n,1}, X_n = 1\right) + \frac{\phi}{2}, & if\ R_{P(\tau_n)+1} = 0 \end{cases} \end{aligned} \tag{136}$$

$$\begin{aligned} &\left(\zeta_n^A | \Theta_{n,1} = \theta_{n,1}, X_n = 1\right) \\ &\quad = \begin{cases} \left(U_{P(\tau_n)+1}\left(t_{P(\tau_n)+1}\right) | \Theta_{n,1} = \theta_{n,1}, X_n = 1\right), & if\ R_{P(\tau_n)+1} = 1 \\ 0, & if\ R_{P(\tau_n)+1} = 0 \end{cases} \end{aligned} \tag{137}$$

And $\forall \theta_{n,1} \in \overline{Q_{n,1,1}}, \forall n = 0, 1, \ldots$:

$$E_R[\zeta_n | \Theta_{n,1} = \theta_{n,1}, X_n = 1] \geq \left(U_{P(\tau_n)+1}\left(t_{P(\tau_n)+1}\right) | \Theta_{n,1} = \theta_{n,1}, X_n = 1\right) + \frac{\phi(1-p)}{2} \tag{138}$$

$$E_R[\zeta_n^A | \Theta_{n,1} = \theta_{n,1}, X_n = 1] = \left(U_{P(\tau_n)+1}\left(t_{P(\tau_n)+1}\right) | \Theta_{n,1} = \theta_{n,1}, X_n = 1\right) p \tag{139}$$

So

$$\begin{aligned} E_R[\zeta_n^A | \Theta_{n,1} = \theta_{n,1}, X_n = 1] \leq E_R[\zeta_n | \Theta_{n,1} = \theta_{n,1}, X_n = 1] p - \frac{\phi p(1-p)}{2} \\ , \forall \theta_{n,1} \in \overline{Q_{n,1,1}}, \forall n = 0, 1, \ldots \end{aligned} \tag{140}$$

Then we have:

$$
\begin{aligned}
&E\left[\zeta_n^A|\Theta_{n,1}\in\overline{Q_{n,1,1}},X_n=1\right]\\
&=\int_{\overline{Q_{n,1,1}}}Pr\left\{\Theta_{n,1}=\theta_{n,1}|\Theta_{n,1}\in\overline{Q_{n,1,1}}\right\}\cdot E_R[\zeta_n^A|\Theta_{n,1}=\theta_{n,1}\in\overline{Q_{n,1,1}},X_n=k]d\theta_{n,1}\\
&\leq\int_{\overline{Q_{n,1,1}}}Pr\left\{\Theta_{n,1}=\theta_{n,1}|\Theta_{n,1}\in\overline{Q_{n,1,1}}\right\}\cdot\left(E_R[\zeta_n|\Theta_{n,1}=\theta_{n,1}\in\overline{Q_{n,1,1}},X_n=k]p-\frac{\phi p\,(1-p)}{2}\right)d\theta_{n,1}
\end{aligned}
\tag{141}
$$

$$
\begin{aligned}
&=\int_{\overline{Q_{n,1,1}}}Pr\left\{\Theta_{n,1}=\theta_{n,1}|\Theta_{n,1}\in\overline{Q_{n,1,1}}\right\}\cdot\left(E_R[\zeta_n|\Theta_{n,1}=\theta_{n,1}\in\overline{Q_{n,1,1}},X_n=k]p\right)d\theta_{n,1}\\
&-\int_{\overline{Q_{n,1,1}}}Pr\left\{\Theta_{n,1}=\theta_{n,1}|\Theta_{n,1}\in\overline{Q_{n,1,1}}\right\}\cdot\left(\frac{\phi p\,(1-p)}{2}\right)d\theta_{n,1}\\
&=E\left[\zeta_n|\Theta_{n,1}\in\overline{Q_{n,1,1}},X_n=1\right]p-\frac{\phi p\,(1-p)}{2}
\end{aligned}
\tag{142}
$$

So for the set $\overline{Q_{n,1,1}}$:

$$
E\left[\zeta_n^A|\Theta_{n,1}\in\overline{Q_{n,1,1}},X_n=1\right]\leq E\left[\zeta_n|\Theta_{n,1}\in\overline{Q_{n,1,1}},X_n=1\right]p-\frac{\phi p\,(1-p)}{2}
\tag{143}
$$

By the Law of Total Expectation, consider the set $Q_{n,1,1}$ and we know that:

$$
\begin{aligned}
&E\left[\zeta_n|\Theta_{n,1}\in Q_{n,1,1},X_n=1\right]\\
&=Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n|\Theta_{n,1}\in\overline{Q_{n,1,1}},X_n=1\right]\\
&+Pr\left\{\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right)|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n|\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right),X_n=1\right]
\end{aligned}
\tag{144}
$$

$$
\begin{aligned}
&E\left[\zeta_n^A|\Theta_{n,1}\in Q_{n,1,1},X_n=1\right]\\
&=Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n^A|\Theta_{n,1}\in\overline{Q_{n,1,1}},X_n=1\right]\\
&+Pr\left\{\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right)|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n^A|\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right),X_n=1\right]
\end{aligned}
\tag{145}
$$

where $Q_{n,1,1}\setminus\overline{Q_{n,1,1}}$ is the set difference of $Q_{n,1,1}$ and $\overline{Q_{n,1,1}}$. The conditional probability $Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}|\Theta_{n,1}\in Q_{n,1,1}\right\}$ can be calculated as follow:

$$
\begin{aligned}
&Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}|\Theta_{n,1}\in Q_{n,1,1}\right\}\\
&=\frac{Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}},\Theta_{n,1}\in Q_{n,1,1}\right\}}{Pr\left\{\Theta_{n,1}\in Q_{n,1,1}\right\}}
\end{aligned}
\tag{146}
$$

$$
=\frac{Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}\right\}}{Pr\left\{\Theta_{n,1}\in Q_{n,1,1}\right\}}
\tag{147}
$$

where the probabilities $Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}\right\}$ and $Pr\left\{\Theta_{n,1}\in Q_{n,1,1}\right\}$ can be derived as follow:

$$
\begin{aligned}
&Pr\left\{\Theta_{n,1} \in Q_{n,1,1}\right\} \\
&= Pr\left\{t_{P(\tau_n)+1} - \tau_n < \frac{\phi}{\mu}\right\} && (148)\\
&= 1 - e^{-\lambda\frac{\phi}{\mu}} && (149)\\
&Pr\left\{\Theta_{n,1} \in \overline{Q_{n,1,1}}\right\} \\
&= Pr\left\{t_{P(\tau_n)+1} - \tau_n < \frac{\phi}{2\mu} \;\&\; t_{P(\tau_n)+2} - t_{P(\tau_n)+1} > \frac{\phi}{2\mu}\right\} && (150)\\
&= Pr\left\{t_{P(\tau_n)+1} - \tau_n < \frac{\phi}{2\mu}\right\} \cdot Pr\left\{t_{P(\tau_n)+2} - t_{P(\tau_n)+1} > \frac{\phi}{2\mu}\right\} && (151)\\
&= \left(1 - e^{-\lambda\frac{\phi}{2\mu}}\right)\left(e^{-\lambda\frac{\phi}{2\mu}}\right) = e^{-\lambda\frac{\phi}{2\mu}} - e^{-\lambda\frac{\phi}{\mu}} && (152)
\end{aligned}
$$

So one can see that $Pr\left\{\Theta_{n,1} \in \overline{Q_{n,1,1}} | \Theta_{n,1} \in Q_{n,1,1}\right\} > 0$, and $Pr\left\{\Theta_{n,1} \in \left(Q_{n,1,1} \setminus \overline{Q_{n,1,1}}\right) | \Theta_{n,1} \in Q_{n,1,1}\right\} > 0$. Equation (131) can be applied to $\forall \theta_{n,1} \in \left(Q_{n,1,1} \setminus \overline{Q_{n,1,1}}\right) \subsetneq Q_{n,1,1}$, so we should have $\forall \theta_{n,1} \in \left(Q_{n,1,1} \setminus \overline{Q_{n,1,1}}\right)$:

$$
\begin{aligned}
E_R[\zeta_n^A | \Theta_{n,1} = \theta_{n,1}, X_n = 1] &= \left(U_{P(\tau_n)+1}\left(t_{P(\tau_n)+1}\right) | \Theta_{n,1} = \theta_{n,1}, X_n = 1\right) \\
&\leq E_R[\zeta_n | \Theta_{n,1} = \theta_{n,1}, X_n = 1]p \qquad (153)\\
\forall \theta_{n,1} \in \left(Q_{n,1,1} \setminus \overline{Q_{n,1,1}}\right), \forall n = 0, 1, \ldots &
\end{aligned}
$$

and consequently:

$$
E\left[\zeta_n^A | \Theta_{n,1} \in \left(Q_{n,1,1} \setminus \overline{Q_{n,1,1}}\right), X_n = 1\right] \leq E\left[\zeta_n | \Theta_{n,1} \in \left(Q_{n,1,1} \setminus \overline{Q_{n,1,1}}\right), X_n = 1\right] p \qquad (154)
$$

By combining this equation with (143), we are able to compare Equation (144) and (145). We have:

$$
\begin{aligned}
&E\left[\zeta_n^A|\Theta_{n,1}\in Q_{n,1,1}, X_n=1\right]\\
&=Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n^A|\Theta_{n,1}\in\overline{Q_{n,1,1}}, X_n=1\right]\\
&+Pr\left\{\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right)|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n^A|\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right), X_n=1\right] \qquad (155)\\
&\le Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n|\Theta_{n,1}\in\overline{Q_{n,1,1}}, X_n=1\right]p\\
&+Pr\left\{\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right)|\Theta_{n,1}\in Q_{n,1,1}\right\}\left(E\left[\zeta_n|\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right), X_n=1\right]p-\frac{\phi p(1-p)}{2}\right) \qquad (156)\\
&=Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n|\Theta_{n,1}\in\overline{Q_{n,1,1}}, X_n=1\right]p\\
&+Pr\left\{\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right)|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n|\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right), X_n=1\right]p\\
&-Pr\left\{\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right)|\Theta_{n,1}\in Q_{n,1,1}\right\}\frac{\phi p(1-p)}{2}\\
&<Pr\left\{\Theta_{n,1}\in\overline{Q_{n,1,1}}|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n|\Theta_{n,1}\in\overline{Q_{n,1,1}}, X_n=1\right]p\\
&+Pr\left\{\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right)|\Theta_{n,1}\in Q_{n,1,1}\right\}E\left[\zeta_n|\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right), X_n=1\right]p \qquad (157)\\
&=E\left[\zeta_n|\Theta_{n,1}=\theta_{n,1}\in Q_{n,1,1}, X_n=1\right]p \qquad (158)
\end{aligned}
$$

Equation(155) is from (145). Equation (156) is from (143). Equation (157) is by removing term $-Pr\left\{\Theta_{n,1}\in\left(Q_{n,1,1}\setminus\overline{Q_{n,1,1}}\right)|\Theta_{n,1}\in Q_{n,1,1}\right\}\frac{\phi p(1-p)}{2}$ which is strictly negative. Then (158) is from (144). So we finally have:

$$E\left[\zeta_n^A|\Theta_{n,1}\in Q_{n,1,1}, X_n=1\right]<E\left[\zeta_n|\Theta_{n,1}\in Q_{n,1,1}, X_n=1\right]p \qquad (159)$$

where (124) directly follows.

We have proved (122), (123) and (124) by now, so Lemma 2 is proved. □

Lemma 2 can be interpreted as follow. In (τ_n,τ_{n+1}), if ζ_n bits of proactive service can all be potentially realized, we should have $E\left[\zeta_n^A\right]=E\left[\zeta_n\right]p$ based on our assumptions on the request processes. This is the case when a transition $X_{n+1}>1$ happens, when every bit of ζ_n is done before the corresponding request arrives. However if a transition $X_{n+1}=1$ happens, the amount of proactive work done in (τ_n,τ_{n+1}) that can potentially be realized is no more than ζ_n, leading to the inequality $E\left[\zeta_n^A\right]\le E\left[\zeta_n\right]p$ in this scenario. An example is shown in (τ_3,τ_4) of Figure 4. The amount of proactive work done in (τ',τ_4) is for request 5 which has arrived but not realized. This part of proactive work will never be realized, so it is not included in ζ_n^A which causes the inequality. Specifically if transitions $X_n=1, X_{n+1}=1$ happen, we proved that strict inequality is achieved.

Intuitively if the transition to the state 1 happens comparably often as all transitions, it will be most likely that $\overline{U}>\overline{U_A}$, based on Lemma 2. Then we proceed to prove Theorem 4.

Proof of Theorem 4: Define $N(t)\triangleq\max\left\{n|J\left(\tau_n^+\right)\le I(t)\right\}$ as the index of the transition where the latest actual request received proactive service. From Proposition 4, we know that the expected time before next transition is finite. So as $t\to\infty$, we know $N(t)\to\infty$ as well. And $\lim_{t\to\infty}\frac{t}{N(t)}=E\left[T_n\right], w.p.1$ where $E\left[T_n\right]$ is a finite constant given system parameters. Recall that we define ζ_n as the amount of proactive work done in (τ_n,τ_{n+1}), and ζ_n^A as the amount of proactive work done for actual requests in (τ_n,τ_{n+1}).

Consider the term $\sum_{i=1}^{I(t)} U_i$. If we rewrite this term from the point of view of transitions, we have:

$$\sum_{i=1}^{I(t)} U_i = \sum_{n=0}^{N(t)} \zeta_n - o(t) \tag{160}$$

where the term $o(t)$ represents the amount of proactive work done in $\left(\tau_{N(t)}, \tau_{N(t)+1}\right)$ for requests which arrive later than $I(t)$. We know $o(t) < \zeta_{N(t)}$ by definition, and we know that $\frac{E[\zeta_n]}{\mu} \leq E[T_n] < \infty, w.p.1$, so we have:

$$\lim_{t\to\infty} \frac{o(t)}{t} = 0, w.p.1 \tag{161}$$

Then we have:

$$\overline{U} = \lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i}{I(t)} \tag{162}$$

$$= \lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \zeta_n - o(t)}{I(t)} \tag{163}$$

$$= \lim_{t\to\infty} \left(\frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_{n+1} = 1)\zeta_n}{N(t)} + \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_{n+1} > 1)\zeta_n}{N(t)} \right) \frac{N(t)}{I(t)} \tag{164}$$

$$= \lim_{t\to\infty} \left(\frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n = 1, X_{n+1} = 1)\zeta_n}{N(t)} + \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n > 1, X_{n+1} = 1)\zeta_n}{N(t)} \right.$$
$$\left. + \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_{n+1} > 1)\zeta_n}{N(t)} \right) \frac{N(t)}{I(t)} \tag{165}$$

where (164) and (165) are by grouping the terms based on transitions. The term $\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_{n+1}=1)}{N(t)}$ represents the limiting fraction of state 1, and the other terms in similar form can be interpreted correspondingly.

Similarly we have:

$$\overline{U_A} = \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+: R_i=1, i\leq I(t)} U_i}{A(t)}$$

$$= \lim_{t\to\infty} \left(\frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_{n+1} = 1)\zeta_n^A}{N(t)} + \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_{n+1} > 1)\zeta_n^A}{N(t)} \right) \frac{N(t)}{A(t)} \tag{166}$$

$$= \lim_{t\to\infty} \left(\frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n = 1, X_{n+1} = 1)\zeta_n^A}{N(t)} + \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n > 1, X_{n+1} = 1)\zeta_n^A}{N(t)} \right.$$
$$\left. + \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_{n+1} > 1)\zeta_n^A}{N(t)} \right) \frac{N(t)}{A(t)} \tag{167}$$

Case 1: If $\phi < U^*$, we know the Markov chain is transient. So we have

$$\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n = 1)}{N(t)} = 0, w.p.1 \tag{168}$$

$$\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n > 1)}{N(t)} = 1, w.p.1 \tag{169}$$

Therefore based on Lemma 2 and Strong Law of Large Numbers, we have:

$$\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)\zeta_n}{N\left(t\right)}=\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)\zeta_n}{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)}{N\left(t\right)} \tag{170}$$

$$=0,w.p.1 \tag{171}$$

$$\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)\zeta_n^A}{N\left(t\right)}=\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)\zeta_n^A}{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)}{N\left(t\right)} \tag{172}$$

$$=0,w.p.1 \tag{173}$$

$$\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)\zeta_n}{N\left(t\right)}=\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)\zeta_n}{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)}{N\left(t\right)}$$

$$=E\left[\zeta_n|X_{n+1}>1\right]\cdot 1 \tag{174}$$

$$=\phi,w.p.1 \tag{175}$$

$$\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)\zeta_n^A}{N\left(t\right)}=\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)\zeta_n^A}{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)}{N\left(t\right)}$$

$$=E\left[\zeta_n^A|X_{n+1}>1\right]\cdot 1 \tag{176}$$

$$=\phi p,w.p.1 \tag{177}$$

And $\lim_{t\to\infty}\frac{A(t)}{N(t)}$ should be the average number of actual arrivals between two consecutive transitions, which converges to $\lambda pE\left[T_n\right]$ by the Law of Large Numbers. So from (164) and (166) we have:

$$\overline{U}=\lim_{t\to\infty}\left(\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)\zeta_n}{N\left(t\right)}+\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)\zeta_n}{N\left(t\right)}\right)\frac{N\left(t\right)}{I\left(t\right)}$$

$$=\frac{\phi}{\lambda E\left[T_n\right]},w.p.1 \tag{178}$$

$$\overline{U_A}=\lim_{t\to\infty}\left(\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)\zeta_n^A}{N\left(t\right)}+\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)\zeta_n^A}{N\left(t\right)}\right)\frac{N\left(t\right)}{A\left(t\right)} \tag{179}$$

$$=\phi p\frac{1}{p\lambda E\left[T_n\right]}=\frac{\phi}{\lambda E\left[T_n\right]},w.p.1 \tag{180}$$

Therefore we have $\overline{U}=\overline{U_A},w.p.1$ when $\phi<U^*$.

Case 2: If $\phi=U^*$, the Markov chain is null recurrent. So we have

$$\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}=1\right)}{N\left(t\right)}=0,w.p.1 \tag{181}$$

$$\lim_{t\to\infty}\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)}{N\left(t\right)}=1,w.p.1 \tag{182}$$

Then the deductions are similar Case 1, so we directly show the conclusions:

$$\overline{U} = \frac{\phi}{\lambda E\left[T_n\right]}, w.p.1 \tag{183}$$

$$\overline{U_A} = \frac{\phi}{\lambda E\left[T_n\right]}, w.p.1 \tag{184}$$

Therefore we have $\overline{U} = \overline{U_A}, w.p.1$ when $\phi = U^*$.

Case 3: If $\phi > U^*$, the Markov chain is positive recurrent. So the Markov chain has a limiting distribution, or steady state probability $\{\pi_k, k = 1, 2, \ldots\}$, where $\pi_k \triangleq Pr\left\{\lim_{n\to\infty} X_n = k\right\}, \forall k \in \mathbb{Z}^+$. Then we have:

$$\begin{aligned}
&\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = 1, X_{n+1} = 1\right)\zeta_n}{N\left(t\right)} \\
&= \lim_{t\to\infty} \left(\frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = 1, X_{n+1} = 1\right)\zeta_n}{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = 1, X_{n+1} = 1\right)} \frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = 1, X_{n+1} = 1\right)}{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = 1\right)} \frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = 1\right)}{N\left(t\right)} \right) && (185) \\
&= E\left[\zeta_n | X_n = 1, X_{n+1} = 1\right] Pr\left\{X_{n+1} = 1 | X_n = 1\right\} \pi_1, w.p.1 && (186) \\
&= \pi_1 \sum_{k=1}^{\infty} p_k^{\phi} E\left[\zeta_n | X_n = 1, X_{n+1} = 1\right], w.p.1 && (187)
\end{aligned}$$

In (185), $\frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n=1, X_{n+1}=1)\zeta_n}{\sum_{n=0}^{N(t)} \mathbb{1}(X_n=1, X_{n+1}=1)}$ is the average of ζ_n between two consecutive transitions where $X_n = 1, X_{n+1} = 1$, which converges to $E\left[\zeta_n | X_n = 1, X_{n+1} = 1\right]$. $\frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n=1, X_{n+1}=1)}{\sum_{n=0}^{N(t)} \mathbb{1}(X_n=1)}$ is the fraction of next transition where $X_{n+1} = 1$ given $X_n = 1$, which converges to transition probability $Pr\left\{X_{n+1} = 1 | X_n = 1\right\}$. $\frac{\sum_{n=0}^{N(t)} \mathbb{1}(X_n=1)}{N(t)}$ is the fraction of state 1, which converges to π_1 in positive recurrent case. Therefore we have (186) based on the Strong Law of Large Numbers. Following similar arguments, we have the following results:

$$\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = k, X_{n+1} = 1\right)\zeta_n}{N\left(t\right)} = \pi_k \left(\sum_{i=k}^{\infty} p_i^{\phi} \right) E\left[\zeta_n | X_n = k, X_{n+1} = 1\right], w.p.1, \forall k > 1 \tag{188}$$

$$\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_{n+1} > 1\right)\zeta_n}{N\left(t\right)} = \left(1 - \pi_1\right)\phi, w.p.1 \tag{189}$$

$$\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = 1, X_{n+1} = 1\right)\zeta_n^A}{N\left(t\right)} = \pi_1 \sum_{i=1}^{\infty} p_i^{\phi} E\left[\zeta_n^A | X_n = 1, X_{n+1} = 1\right], w.p.1 \tag{190}$$

$$\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_n = k, X_{n+1} = 1\right)\zeta_n^A}{N\left(t\right)} = \pi_k \left(\sum_{i=k}^{\infty} p_i^{\phi} \right) E\left[\zeta_n^A | X_n = k, X_{n+1} = 1\right], w.p.1, \forall k > 1 \tag{191}$$

$$\lim_{t\to\infty} \frac{\sum_{n=0}^{N(t)} \mathbb{1}\left(X_{n+1} > 1\right)\zeta_n^A}{N\left(t\right)} = \left(1 - \pi_1\right)\phi p, w.p.1 \tag{192}$$

based on Lemma 2 and the Strong Law of Large Numbers. So we have:

$$\begin{aligned}\overline{U} &= \lim_{t\to\infty}\left(\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_n=1,X_{n+1}=1\right)\zeta_n}{N(t)}+\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_n=1,X_{n+1}=1\right)\zeta_n}{N(t)}\right.\\&\left.+\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)\zeta_n}{N(t)}\right)\frac{N(t)}{I(t)} \qquad (193)\\&= \frac{\pi_1\sum_{k=1}^{\infty}p_k^{\phi}E\left[\zeta_n|X_n=1,X_{n+1}=1\right]}{\lambda E\left[T_n\right]}+\frac{\sum_{k=2}^{\infty}\pi_k\left(\sum_{i=k}^{\infty}p_i\right)E\left[\zeta_n|X_n=k,X_{n+1}=1\right]}{\lambda E\left[T_n\right]}\\&+\frac{\left(1-\pi_1\right)\phi}{\lambda E\left[T_n\right]}\\\overline{U_A} &= \lim_{t\to\infty}\left(\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_n=1,X_{n+1}=1\right)\zeta_n^A}{N(t)}+\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_n=1,X_{n+1}=1\right)\zeta_n^A}{N(t)}\right.\\&\left.+\frac{\sum_{n=0}^{N(t)}\mathbb{1}\left(X_{n+1}>1\right)\zeta_n^A}{N(t)}\right)\frac{N(t)}{A(t)} \qquad (194)\\&= \frac{\pi_1\sum_{k=1}^{\infty}p_k^{\phi}E\left[\zeta_n^A|X_n=1,X_{n+1}=1\right]}{p\lambda E\left[T_n\right]}+\frac{\sum_{k=2}^{\infty}\pi_k\left(\sum_{i=k}^{\infty}p_i\right)E\left[\zeta_n^A|X_n=k,X_{n+1}=1\right]}{p\lambda E\left[T_n\right]} \qquad (195)\\&+\frac{\left(1-\pi_1\right)\phi p}{p\lambda E\left[T_n\right]}\\&< \frac{\pi_1\sum_{k=1}^{\infty}p_k^{\phi}E\left[\zeta_n|X_n=1,X_{n+1}=1\right]}{\lambda E\left[T_n\right]}+\frac{\sum_{k=2}^{\infty}\pi_k\left(\sum_{i=k}^{\infty}p_i\right)E\left[\zeta_n|X_n=k,X_{n+1}=1\right]}{\lambda E\left[T_n\right]}\\&+\frac{\left(1-\pi_1\right)\phi}{\lambda E\left[T_n\right]} \qquad (196)\\&= \overline{U} \qquad (197)\end{aligned}$$

the strict inequality in (196) is from (124) of Lemma 2. Therefore we have $\overline{U_A}<\overline{U}, w.p.1$ if $\phi>U^*$.

By summarizing Cases 1, 2 and 3, the threshold-based strategy Ψ_P^{ϕ} satisfies Property 2 if and only if $\phi\le U^*$.

J PROOF OF COROLLARY 3

The UNIFORM strategy satisfies both Property 1 and Property 2 by Theorem 2, therefore Corollary 1 can be applied. So we have :

$$\overline{U}=U^*, w.p.1 \qquad (198)$$

Then we select such sample paths where $\overline{U}=U^*$ is satisfied. For every sample path of this set, we assume that $\forall\epsilon>0, \exists\delta>0$ such that:

$$\lim_{t\to\infty}\frac{1}{I(t)}\sum_{i=1}^{I(t)}\mathbb{1}\left(U_i<U^*-\epsilon\right)=\delta \qquad (199)$$

Define sets $H_\epsilon^- (t) = \{i : U_i < U^* - \epsilon, i \le I(t), i \in \mathbb{Z}^+\}$ and $H_\epsilon^+ (t) = \{i : U_i \ge U^* - \epsilon, i \le I(t), i \in \mathbb{Z}^+\}$, then we have:

$$\begin{aligned}
\overline{U} &= \lim_{t\to\infty} \frac{\sum_{i=1}^{I(t)} U_i}{I(t)} \\
&= \lim_{t\to\infty} \sum_{i\in H_\epsilon^-(t)} \frac{U_i}{I(t)} + \lim_{t\to\infty} \sum_{i\in H_\epsilon^+(t)} \frac{U_i}{I(t)} && (200)\\
&= \lim_{t\to\infty} \sum_{i\in H_\epsilon^-(t)} \frac{|H_\epsilon^-(t)|}{I(t)} \frac{U_i}{|H_\epsilon^-(t)|} + \lim_{t\to\infty} \sum_{i\in H_\epsilon^+(t)} \frac{|H_\epsilon^+(t)|}{I(t)} \frac{U_i}{|H_\epsilon^+(t)|} && (201)\\
&\le \lim_{t\to\infty} \sum_{i\in H_\epsilon^-(t)} \frac{|H_\epsilon^-(t)|}{I(t)} \frac{U^* - \epsilon}{|H_\epsilon^-(t)|} + \lim_{t\to\infty} \sum_{i\in H_\epsilon^+(t)} \frac{|H_\epsilon^+(t)|}{I(t)} \frac{U^*}{|H_\epsilon^+(t)|} && (202)\\
&= \delta(U^* - \epsilon) + (1-\delta) U^* && (203)\\
&= U^* - \delta\epsilon \\
&< U^* && (204)
\end{aligned}$$

The reason of Equation (200) is by grouping all U_i into two sets according to if $U_i < U^* - \epsilon$ or not. By replacing all U_i by $U^* - \epsilon$ in the group of $H_\epsilon^+ (t)$ where $U_i < U^* - \epsilon$ and replacing all U_i by U^* in the group of $H_\epsilon^+ (t)$ where $U_i < U^*$, we get (202) from (201). We get (203) from (202) based on our assumption in (199).

However, (204) contradicts the way we selected the sample path. Therefore, for this set of sample paths, we have $\forall \epsilon > 0$:

$$\lim_{t\to\infty} \frac{1}{I(t)} \sum_{i=1}^{I(t)} \mathbb{1}(U_i < U^* - \epsilon) = 0 \tag{205}$$

Therefore we have our conclusions for all the possible sample paths:

$$\lim_{t\to\infty} \frac{1}{I(t)} \sum_{i=1}^{I(t)} \mathbb{1}(U_i = U^*) = 1, w.p.1 \tag{206}$$

K PROOF OF COROLLARY 4

We use the Pollaczek-Khinchine formula in the analysis of M/G/1 queue in [5] to conduct this analysis. The average unfinished work in number of bits in the system by time t, which is defined as $v(t)$, can be formulated as:

$$v(t) = \frac{\sum_{i\in\mathbb{Z}^+: R_i=1, i\le I(t)} \left(S_i W_i + \frac{1}{2} S_i \left(\frac{S_i}{\mu} \right) \right)}{t} \tag{207}$$

The corresponding terms are as shown in Figure 3. S_i is the reactive work of actual request i, and W_i is the waiting time of the reactive part of request i when it starts to be transmitted. Define $v = \lim_{t\to\infty} v(t)$ and take the limit of $t \to \infty$ of (207):

$$\begin{aligned}
v &= \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+: R_i=1, i\le I(t)} \left(S_i W_i + \frac{1}{2} S_i \left(\frac{S_i}{\mu} \right) \right)}{t} && (208)\\
&= \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+: R_i=1, i\le I(t)} S_i W_i}{t} + \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+: R_i=1, i\le I(t)} S_i^2}{2t\mu} && (209)
\end{aligned}$$

Consider the term $\sum_{i\in\mathbb{Z}^+:R_i=1,\,i\le I(t)} S_iW_i$ in Equation (209), we have:

$$\sum_{i\in\mathbb{Z}^+:R_i=1,\,i\le I(t)} S_iW_i = \sum_{i\in\mathbb{Z}^+:R_i=1,\,S_i>S^*,\,i\le I(t)} S_iW_i + \sum_{i\in\mathbb{Z}^+:R_i=1,\,S_i=S^*,\,i\le I(t)} S_iW_i \tag{210}$$

Define sets $H_{S^*}(t)=\{i\in\mathbb{Z}^+ : R_i=1, S_i=S^*, i\le I(t)\}$ and $H^+_{S^*}(t)=\{i\in\mathbb{Z}^+ : R_i=1, S_i>S^*, i\le I(t)\}$, then divide both sides with $I(t)$:

$$\frac{\sum_{i\in\mathbb{Z}^+:R_i=1,\,i\le I(t)} S_iW_i}{I(t)} = \frac{\sum_{i\in H^+_{S^*}(t)} S_iW_i}{I(t)} + \frac{\sum_{i\in H_{S^*}(t)} S_iW_i}{I(t)} \tag{211}$$

$$= \frac{|H^+_{S^*}(t)|}{I(t)}\frac{\sum_{i\in H^+_{S^*}(t)} S_iW_i}{|H^+_{S^*}(t)|} + \frac{|H_{S^*}(t)|}{I(t)}\frac{\sum_{i\in H_{S^*}(t)} S_iW_i}{|H_{S^*}(t)|} \tag{212}$$

Take limit of $t\to\infty$ on both sides and we can get:

$$\lim_{t\to\infty}\frac{\sum_{i\in\mathbb{Z}^+:R_i=1,\,i\le I(t)} S_iW_i}{I(t)} = \lim_{t\to\infty}\frac{|H^+_{S^*}(t)|}{I(t)}\frac{\sum_{i\in H^+_{S^*}(t)} S_iW_i}{|H^+_{S^*}(t)|} + \lim_{t\to\infty}\frac{|H_{S^*}(t)|}{I(t)}\frac{\sum_{i\in H_{S^*}(t)} S_iW_i}{|H_{S^*}(t)|} \tag{213}$$

$$= \lim_{t\to\infty}\frac{1}{I(t)}\sum_{i=1}^{I(t)}\mathbb{1}(S_i>S^*)\frac{\sum_{i\in H^+_{S^*}(t)} S_iW_i}{|H^+_{S^*}(t)|} + \lim_{t\to\infty}\frac{1}{I(t)}\sum_{i=1}^{I(t)}\mathbb{1}(S_i=S^*)\frac{\sum_{i\in H_{S^*}(t)} S_iW_i}{|H_{S^*}(t)|} \tag{214}$$

Because the network scenario we are considering is $\lambda ps<\mu$, so all the W_i are bounded w.p.1. By Corollary 3 we have:

$$\lim_{t\to\infty}\frac{\sum_{i\in\mathbb{Z}^+:R_i=1,\,i\le I(t)} S_iW_i}{I(t)} = \lim_{t\to\infty} 0\cdot\frac{\sum_{i\in H^+_{S^*}(t)} S_iW_i}{|H^+_{S^*}(t)|} + \lim_{t\to\infty} 1\cdot\frac{\sum_{i\in H_{S^*}(t)} S_iW_i}{|H_{S^*}(t)|}, w.p.1 \tag{215}$$

$$= (S^*)\lim_{t\to\infty}\frac{\sum_{i\in H_{S^*}(t)} W_i}{|H_{S^*}(t)|}, w.p.1 \tag{216}$$

$$= (S^*)\lim_{t\to\infty}\frac{\sum_{i\in H_{S^*}(t)} W_i}{A(t)}, w.p.1 \tag{217}$$

Because of Corollary 3, we have (215), and we have $\lim_{t\to\infty}\frac{A(t)}{t}=\lim_{t\to\infty}\frac{|H_{S^*}(t)|}{t}$ w.p.1 for (217). The reason for (216) is by the definition of $H_{S^*}(t)$ so we can replace all S_i with S^*. Define $w\triangleq\lim_{t\to\infty}\frac{\sum_{i\in\mathbb{Z}^+:R_i=1,\,i\le I(t)} W_i}{A(t)}$ and we have:

$$w\triangleq\lim_{t\to\infty}\frac{\sum_{i\in\mathbb{Z}^+:R_i=1,\,i\le I(t)} W_i}{A(t)} = \lim_{t\to\infty}\frac{\sum_{i\in H_{S^*}(t)} W_i}{A(t)}, w.p.1 \tag{218}$$

So (209) can be transformed as follow:

$$v = \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} S_i W_i}{A(t)} \frac{A(t)}{t} + \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} S_i^2}{A(t)} \frac{A(t)}{2t\mu}, w.p.1 \tag{219}$$

$$= (S^*)\, w\lambda p + \frac{(S^*)^2 \lambda p}{2\mu}, w.p.1 \tag{220}$$

Because of the important property of a Poisson process, namely the Poisson-Arrivals-See-Time-Averages (PASTA)[5], we have $\frac{v}{\mu} = w, w.p.1$, and we have:

$$w = \frac{(S^*)^2 \lambda p}{2\mu\,(\mu - (S^*)\,\lambda p)}, w.p.1 \tag{221}$$

And for limiting average delay we have:

$$\overline{D} = \frac{(S^*)^2 \lambda p}{2\mu\,(\mu - (S^*)\,\lambda p)} + \frac{S^*}{\mu}, w.p.1 \tag{222}$$

The following calculations can be done by replacing S^* with $\frac{\lambda s - \mu}{\lambda(1-p)}$.

L PROOF OF COROLLARY 5

Following Equation (222) and Corollary 3, the delay for UNIFORM strategy is:

$$\overline{D}_U = \frac{S^{*2} \lambda p}{2\mu\,(\mu - S^* \lambda p)} + \frac{S^*}{\mu}, w.p.1 \tag{223}$$

With the EDF strategy, we need to consider Equation (209). According to the design of the EDF strategy, the actual requests in the same busy period have the following relationship. If an actual request i is proactively served, no matter partially or fully, the corresponding waiting time satisfies $W_i = 0$ because all the previous potential requests have either been realized or have been fully proactively served. So we have the following results:

$$S_i = s \Rightarrow W_i \ge 0; S_i < s \Rightarrow W_i = 0, \forall i \in \mathbb{Z}^+ \tag{224}$$

So by reorganizing Equation (209):

$$v = \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} S_i W_i}{t} + \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} S_i^2}{2t\mu}$$

$$= \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, S_i=s, i\le I(t)} S_i W_i}{t}$$

$$+ \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, S_i<s, i\le I(t)} S_i W_i}{t}$$

$$+ \lim_{t\to\infty} \frac{1}{2\mu t} \sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} Si^2 \tag{225}$$

$$= \lim_{t\to\infty} \frac{1}{t} \sum_{i\in\mathbb{Z}^+:R_i=1, S_i=s, i\le I(t)} S_i W_i \tag{226}$$

$$+ \lim_{t\to\infty} \frac{1}{t} \sum_{i\in\mathbb{Z}^+:R_i=1, S_i<s, i\le I(t)} S_i W_i \tag{227}$$

$$+ \lim_{t\to\infty} \frac{1}{2\mu t} \sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} S_i^2 \tag{228}$$

Equation (225) is by splitting the S_iW_i terms into two groups according to whether $S_i < s$ or not. Then in (226), we use s to replace all the S_i where $i \in \{i : R_i = 1, S_i = s, i = 1, \ldots, I(t)\}$ because of (224). Also because of (224) we have $W_i = 0, \forall i \in \{i : R_i = 1, S_i < s, i = 1, \ldots, I(t)\}$. So it would not affect the results if we replace all S_i with s in this group in (227). Combine the terms in (226) and (227) and we have:

$$v = \lim_{t\to\infty} \frac{s}{t}\left(\sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} W_i\right) + \lim_{t\to\infty} \frac{1}{2\mu t} \sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} S_i^2 \tag{229}$$

$$= \lim_{t\to\infty} s\frac{A(t)}{t}\frac{\left(\sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} W_i\right)}{A(t)} + \lim_{t\to\infty} \frac{1}{2\mu}\frac{A(t)}{t}\frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} S_i^2}{A(t)} \tag{230}$$

$$= \lambda p s w_E + \frac{\lambda p}{2\mu}\overline{S_E^2}, w.p.1 \tag{231}$$

where $w_E \triangleq \lim_{t\to\infty} \frac{\sum_{i\in\mathbb{Z}^+:R_i=1, i\le I(t)} W_i}{A(t)}$ is the limiting average of waiting time for each actual request under the EDF strategy, and S_E is the reactive work of requests under the EDF strategy.

Also due to PASTA, we have:

$$w_E = \frac{\lambda p \overline{S_E^2}}{2\mu(\mu - \lambda p s)}, w.p.1 \tag{232}$$

Notice here, s is the original object size without any proactive work. We have $\overline{S_E} > S^*$ due to Theorem 4 and Corollary 1, then $\overline{S_E^2} \ge \left(\overline{S_E}\right)^2 > S^{*2}$, so we have:

$$\begin{aligned}\overline{D}_E &= \frac{\lambda p \overline{S_E^2}}{2(\mu^2 - \mu\lambda p s)} + \frac{\overline{S_E}}{\mu}, w.p.1\\ &\ge \frac{\lambda p S^{*2}}{2(\mu^2 - \mu\lambda p S^*)} + \frac{S^*}{\mu}, w.p.1\\ &= \overline{D}_U\end{aligned} \tag{233}$$

where equality holds if and only if $p = 0$.

Received November 2018; revised December 2018; accepted January 2019